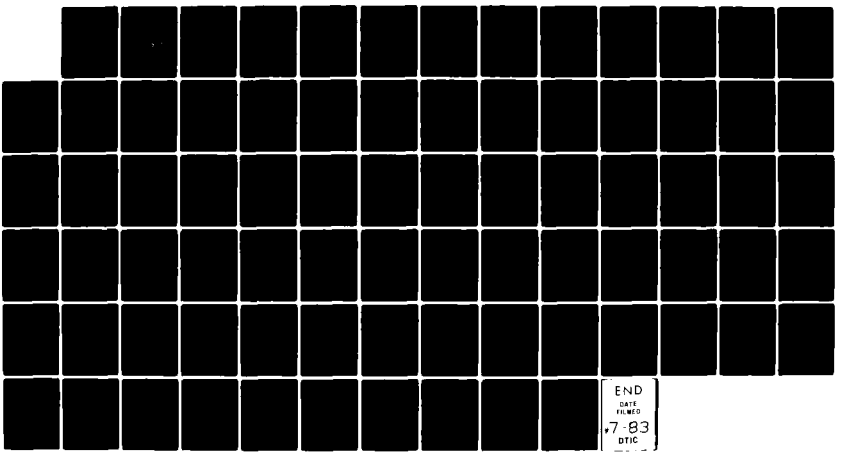
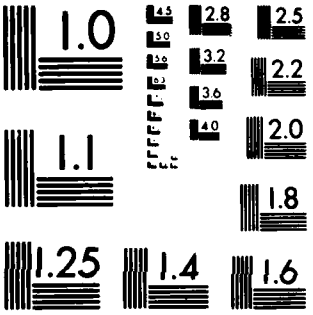
AD-A129 150 BLOCK ITERATIVE METHODS FOR ELLIPTIC FINITE ELEMENT 1/1
EQUATIONS(U) WISCONSIN UNIV-MADISON DEPT OF MATHEMATICS
S V PARTER ET AL. MAR 83 AFOSR-TR-83-0477 W-7405-ENG-36
UNCLASSIFIED F/G 12/1 NL





UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(2)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 83-0477	2. GOVT ACCESSION NO. AD-A129150	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) BLOCK ITERATIVE METHODS FOR ELLIPTIC FINITE ELEMENT EQUATIONS		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Seymour V. Parter and Michael Steuerwalt*		8. CONTRACT OR GRANT NUMBER(s) AFOSR-82-0275
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Mathematics University of Wisconsin Madison WI 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS PE61102F; 2304/A3
11. CONTROLLING OFFICE NAME AND ADDRESS Mathematical & Information Sciences Directorate Air Force Office of Scientific Research Bolling AFB DC 20332		12. REPORT DATE MAR 83 13. NUMBER OF PAGES 75
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

DTIC ELECTE JUN 10 1983 S D A

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
*Michael Steuerwalt is with the University of California, Los Alamos National Laboratory, Los Alamos NM 87545.

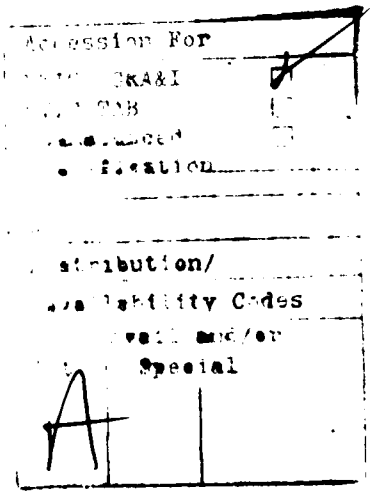
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Direct iterative methods for solving the linear system AU = Y split A into a difference M-N. By viewing N as a weak multiplication operator, the authors determine the convergence rates of block direct iterative methods for solving the system of equations that arises in the finite element approximation of an elliptic boundary value problem. The authors illustrate the theory with an analysis of second order Dirichlet problems in the unit square, using Hermite cube finite element spaces. However, the method of analysis extends to general elliptic boundary value problems of order 2m on bounded domains (CONTINUED)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ITEM #20, CONTINUED: in d space dimensions, and to a broad class of finite element spaces.



AFOSR-TR- 83-0477

BLOCK ITERATIVE METHODS FOR ELLIPTIC FINITE ELEMENT EQUATIONS[1]

Seymour V. Parter[2] *and Michael Steuerwalt*[3]

ABSTRACT

Direct iterative methods for solving the linear system $AU = Y$ split A into a difference $M-N$. By viewing N as a weak multiplication operator, we determine the convergence rates of block direct iterative methods for solving the system of equations that arises in the finite element approximation of an elliptic boundary value problem. We illustrate the theory with an analysis of second order Dirichlet problems in the unit square, using Hermite cubic finite element spaces. However, the method of analysis extends to general elliptic boundary value problems of order $2m$ on bounded domains in d space dimensions, and to a broad class of finite element spaces.

(1) This work was supported by the U.S. Department of Energy under Contract W-7405-Eng-36, and by the Air Force Office of Scientific Research under Contract AFOSR-82-0275.

(2) Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

(3) University of California, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

1. Introduction.

Discrete approximations of linear elliptic partial differential equations lead to a linear system of algebraic equations

$$AU = Y, \tag{1.1}$$

in which the matrix A represents a discretization of the partial differential operator and U is a discrete approximation of the true solution. Typically this is a large, sparse algebraic system: on a mesh of size h in a d-dimensional region, U has $O(h^{-d})$ components, and A has only a few times that many nonzero elements. The development of computers made practical the solution of such systems. Hardware limitations and a desire to solve multidimensional problems, together with the size and sparseness of the system, combined to stimulate the development of direct iterative methods for solving (1.1). Elliptic difference equations, which lead to big systems (1.1) partly because the standard finite difference schemes have $O(h^2)$ accuracy, received special attention: see [8], [22], [1], [12], [20], and [13]. But the development of finite element methods -- particularly higher order accurate methods on irregular meshes -- and of direct factorization methods suitable for finite element systems (1.1) (see e.g. [5], [24], [2], [19], [4], [6], [17], [9]), together with the discovery of fast factorization methods for nice elliptic difference equations (see [15] for some references), combined to lessen interest in iterative methods.

Nevertheless, iterative methods for finite element equations have received some attention. Fix and Larsen [7] and Varga [21] studied the convergence of the successive overrelaxation (SOR) method, based on point and k-line block splittings of the finite element matrix A, for self-adjoint elliptic problems of order $2m$. They showed for such problems that there are choices of the relaxation parameter ω for which the spectral radius ρ_ω satisfies the inequality

$\rho_\omega \leq 1 - Kh^m$; when $\omega = 1$, which is the Gauss-Seidel method, the corresponding inequality is $\rho_{GS} \leq 1 - Kh^{2m}$. Each inequality is what one would expect for finite difference approximations. But in neither instance could they determine the constant K.

In [16] Rice experimentally compared direct factorization methods to point SOR methods for Hermite cubic finite element approximations of some second order elliptic problems. He concluded that point SOR and Jacobi conjugate gradient iterative methods are more efficient than Gaussian elimination when the approximation is sufficiently accurate. For the problems he considered, "sufficient accuracy" is a surprisingly coarse 0.1%.

A direct (or cyclic) iterative scheme splits the matrix A into the difference

$$A = M - N, \tag{1.2}$$

and generates a sequence $\{U^{(\nu)}\}$ according to

$$MU^{(\nu)} = NU^{(\nu-1)} + Y. \tag{1.3}$$

Convergence of the sequence is governed by the spectral radius ρ of $M^{-1}N$: $\{U^{(\nu)}\}$ converges to the solution of (1.1) for any $U^{(0)}$ iff $\rho < 1$, and smaller ρ implies faster convergence. To determine the convergence rate of (1.3) therefore requires not only that we establish estimates like

$$\rho \approx 1 - Kh^p,$$

but also that we determine p and K.

In [13] one of us (Parter) developed a general approach for estimating the rates of convergence of the classical iterative schemes -- Jacobi, Gauss-Seidel, and SOR -- for self-adjoint elliptic finite difference problems. In [15] we

simplified the presentation and extended the method of analysis to parabolic problems and to nonself-adjoint elliptic finite difference problems. The key to the method is that the matrix N looks like a weak multiplication operator: there is a function q for which $(NU,V) \approx (qU,V)$. In this work we employ the same basic approach to deal with finite element equations arising from elliptic problems, even problems that are not self-adjoint. However, the analysis of (1.3) for finite element equations requires several new ideas.

The theory of [13] and [15] asks that the splitting (1.2) satisfy four basic properties. To verify the third (A.3 in this paper, A.4 in [15]), which asserts that N behaves properly, can be a little complicated, even in the finite difference case. For the finite element case it appears to be very difficult. Part of the difficulty stems from the fact that finite element methods involve derivatives as well as function values. For example, in a second order elliptic problem the finite element method based on tensor products of Hermite quintic splines will involve several derivatives beyond the first. These derivatives appear in the elements of N. Nevertheless, the finite element method only yields H^1 estimates – that is, L^2 estimates on the approximate solution and its first derivatives. In [3] Boley and Parter studied a finite element approximation of a simple one-dimensional problem. Their treatment of derivative terms cannot be extended to multidimensional problems.

Sections 6 through 8 discuss the model problem that seeks u satisfying Dirichlet boundary conditions and the equation

$$Lu := -[(au_x)_x + (bu_x)_y + (bu_y)_x + (cu_y)_y] + d_1u_x + d_2u_y + d_0u = f \quad (1.4)$$

in the unit square Ω, with $d_0(x,y) \geq 0$. The finite element subspaces S_n are tensor products of Hermite cubic splines. We consider both k-line iterative methods and the point Gauss-Seidel method. In these cases we find that one

need consider only the function values (as in the finite difference case) -- that is, we can ignore certain derivative terms. Thus the necessary calculations are similar to those carried out in our earlier work [15]. In particular, the spectral radius $\rho_J(k)$ of the k-line block Jacobi iterative method is given by

$$\rho_J(k) \approx 1 - \frac{5k}{12}\Gamma_0 \Delta y^2 \tag{1.5}$$

asymptotically as $\Delta y \to 0$. Here Γ_0 is the minimal eigenvalue of the elliptic eigenvalue problem

$$L\varphi = \lambda c(x,y)\varphi \quad \text{in } \Omega, \quad \varphi = 0 \quad \text{on } \partial\Omega. \tag{1.6}$$

Because this iterative scheme satisfies block property A, the corresponding spectral radius $\rho_\omega(k)$ for the successive overrelaxation (SOR) k-line method with relaxation parameter ω is fixed by the equation

$$(\rho_\omega + \omega - 1)^2 = \omega^2 \rho_J^2 \rho_\omega.$$

Thus the Jacobi spectral radius determines the smallest SOR spectral radius ρ_b, its corresponding ω_b, and the Gauss-Seidel spectral radius ρ_{GS}:

$$\begin{gathered} \rho_{GS} := \rho_J^2, \\ \omega_b := \frac{2}{1 + \sqrt{1 - \rho_J^2}}, \quad \rho_b := \omega_b - 1 \end{gathered} \tag{1.7}$$

(see [1], [20, chapter 4], [23]). It is interesting to compare the estimate (1.5) with our earlier estimates in [12], [13], and [15] for a (particular) finite difference approximation. In that case we have

$$\rho_J(k) \approx 1 - \frac{k}{2}\Gamma_0 \Delta y^2 \tag{1.8}$$

with the same Γ_0!

While the detailed analysis leading to these results is carried out only for the model problem, it is clear that these ideas apply under much more general circumstances. For example, this analysis is easily extended to those cases where the finite element subspaces are "nodal" finite element subspaces (see [18] or [19]) and the block splitting is based on a reasonable geometric choice of blocks. The region Ω may be any smooth region in R^d, while the elliptic operator may be any strongly elliptic operator of order $2m$. In section 9 we comment further on the generality of the analysis contained herein.

Sections 2 through 5 are concerned with the general approach and develop the basic theory. In section 2 we describe the general class of finite element approximations to elliptic boundary value problems that can be written in a weak form. We also describe the related algebraic problems (1.1). In section 3 we recall the classical iterative methods. In section 4 we extend our earlier thoretical work for finite difference approximations to the finite element setting. The basic hypothesis is assumption A.3:

There are a constant q_0 and a function $q \in C^1(\bar{\Omega})$ with

$$q(x) \geq q_0 > 0 \quad (x \in \bar{\Omega}),$$

and a function $\eta(t)$, defined for $t \geq 1$ and satisfying

$$\eta(t) \to 0 \quad \text{as } t \to \infty,$$

such that for every u and v in S_n we have

$$h^{2m}\hat{V}^{*}N\hat{U} = \int qu\bar{v}\,dx + \varepsilon_n(u,v),$$

where

$$|\varepsilon_n(u,v)| \le \eta(n)[\|u\|_1^2 + \|v\|_1^2 + \|u\|_1 + \|v\|_1].$$

While this may appear to be an unusual condition, we believe that it is satisfied by most natural splittings. The basis for this belief for finite differences is described in [15, section 9]. For finite elements, our belief is grounded in that discussion and the fact that we can ignore derivative terms to estimate q. As we shall see later, the exact form of the bounds on the error term $\varepsilon_n(u,v)$ can be exploited to give some interesting results.

In section 5 a new convergence theorem is proven. Loosely speaking, if (i) the subspaces S_n satisfy certain inverse inequalities, (ii) there is a particular bound on ε_n, and (iii) there is an eigenpair $(\lambda, \hat{U})$ associated with the spectral radius ρ so that $|\lambda| = \rho$ and

$$\mathrm{Re}\,(\hat{U}^{*}N\hat{U}) \ge 0, \tag{1.9}$$

then for small h the method is convergent and we can estimate the asymptotic form of ρ. While it is not at all obvious that one should expect (1.9) to hold in the generality of the finite element equations and for nonself-adjoint problems, nevertheless condition (1.9) is always true for block Jacobi schemes that have block property A.

Finally, in Section 9 we discuss the significance of the results.

For the reader's convenience, we collect some notation in the rest of this section.

Let Ω be a bounded domain in $\mathbb{R}^d$. If u and v are in $L^2(\Omega)$, their inner

product is

$$(u,v)_\Omega := \int_\Omega u(x)\overline{v(x)}\,dx.$$

The corresponding norm is denoted by

$$\|u\|_{0;\Omega} := \sqrt{(u,u)_\Omega}.$$

Customarily we write (u,v) and $\|u\|_0$ when the set Ω is clear from the context.

Let $u:\Omega\to\mathbf{R}$. We denote the partial derivative of u with respect to x_i by $D_i u := \frac{\partial u}{\partial x_i}$. Conventionally, if $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_d)$ is a d-tuple of nonnegative integers, then we set $|\alpha| := \alpha_1+\alpha_2+\cdots+\alpha_d$, and by $D^\alpha u$ we mean $D_1^{\alpha_1}\cdots D_d^{\alpha_d}u$.

By $H^m(\Omega)$ we mean the set of functions u that together with all their partial derivatives up to order m are in $L^2(\Omega)$. Symbolically, we have

$$H^m(\Omega) := \{u \in L^2(\Omega) : D^\alpha u \in L^2(\Omega) \text{ for } 0 \le |\alpha| \le m\}.$$

$H^m(\Omega)$ is a Hilbert space with norm defined by

$$\|u\|_m^2 := \sum_{|\alpha|\le m} (D^\alpha u, D^\alpha u).$$

It will be convenient to define the seminorms $|\cdot|_j$ on $H^m(\Omega)$ for $0 \le j \le m$ by

$$|u|_j^2 := \sum_{|\alpha|=j} (D^\alpha u, D^\alpha u). \tag{1.10}$$

Observe that $|\cdot|_0$ is the L^2 norm $\|\cdot\|_0$, and $\|u\|_m^2 = \sum_{0\le j\le m} |u|_j^2$.

Throughout the text, C and K denote generic constants. Constants of more than local importance are numbered.

We are indebted to Carl deBoor and Louis Nirenberg for useful discussions.

2. The problem.

Let Ω be a smooth, bounded domain in $\mathbf{R}^d$, and let L be the linear elliptic operator of order $2m$ defined by

$$Lu := \sum_{|\alpha|,|\beta|\le m} (-1)^{|\alpha|} D^{\alpha}(a_{\alpha,\beta}(x) D^{\beta} u). \tag{2.1}$$

We consider the boundary value problem

$$Lu = f \quad \text{in } \Omega, \quad b_j u = 0 \quad \text{on } \partial\Omega \quad (0 \le j \le m-1), \tag{2.2}$$

where the boundary operators b_j are linear and independent. We assume that the problem (2.2) is equivalent to the following "weak" formulation: there is a subspace $\tilde{H}^m$ of $H^m(\Omega)$ and we seek $u \in \tilde{H}^m$ such that

$$B(u,v) = F(v) \quad \text{for all } v \in \tilde{H}^m, \tag{2.3}$$

where

$$B(u,v) := \int_{\Omega} b(u,v)\, dx \tag{2.4a}$$

with

$$b(u,v) := \sum_{|\alpha|,|\beta|\le m} a_{\alpha,\beta}(x) D^{\beta} u D^{\alpha} \bar{v}, \tag{2.4b}$$

and

$$F(v) := \int_{\Omega} f(x) \overline{v(x)}\, dx. \tag{2.4c}$$

Note that this formulation of the problem is in effect a statement about the nature of the boundary conditions of (2.2).

We also assume that the form $B(u,v)$ is continuous and coercive on $\tilde{H}^m$. That is, we assume there are positive constants K_1 and K_0 such that for all u and v in $\tilde{H}^m$ we have

$$\begin{aligned} |B(u,v)| &\le K_1 \|u\|_m \|v\|_m \quad \text{(continuity)}, \\ \text{Re } B(u,u) &\ge K_0 \|u\|_m^2 \quad \text{(coercivity)}. \end{aligned} \tag{2.5}$$

A simple computation shows for any $u \in \tilde{H}^m$ that

$$|\text{Im } B(u,u)| \le \frac{K_1}{K_0} |\text{Re } B(u,u)|.$$

This inequality is called the *angle-bounded property* of B. A finite element approach to the numerical solution of this problem is given by a sequence $\{S_n\}$ of finite-dimensional subspaces of $\tilde{H}^m$, satisfying

$$\dim(S_n) = n, \tag{2.6}$$

and the solution $u_n \in S_n$ of the finite-dimensional discrete problem

$$B(u_n, v_n) = F(v_n) \quad \text{for all } v_n \in S_n. \tag{2.7}$$

With each S_n we associate a basis $\{\varphi_j\}_{j=1}^n$, and a positive constant $h = h_n$, the "mesh size"; we suppose that $h_n \to 0$ as $n \to \infty$.

The problem (2.7) is brought into computable form by setting

$$a_{i,j} := B(\varphi_j, \varphi_i) \quad (1 \le i, j \le n). \tag{2.8}$$

Problem (2.7) now takes the form: find

$$u_n := \sum_{j=1}^{n} U_j \varphi_j \quad \text{in } S_n. \tag{2.9a}$$

where the vector

$$\hat{U}_n := (U_1, \cdots, U_n)^t \tag{2.9b}$$

corresponding to the function u_n satisfies

$$\sum_{j=1}^{n} a_{i,j} U_j = \int_\Omega f \, \overline{\varphi}_i \, dx = F(\varphi_i) =: F_i \quad (1 \le i \le n). \tag{2.9c}$$

Thus, if A is the $n \times n$ matrix and $\hat{F}$ is the n-vector

$$A := (a_{i,j}), \tag{2.10a}$$

$$\hat{F} := (F_1, \cdots, F_n)^t, \tag{2.10b}$$

then (2.7) reduces to the problem of finding $\hat{U}$ that solves

$$A\hat{U} = \hat{F}. \tag{2.10c}$$

The matrix A is called the *problem matrix*. Another matrix of interest is the *mass matrix* Q given by

$$Q_{i,j} := \int_\Omega \varphi_j \overline{\varphi_i} \, dx. \tag{2.11}$$

If $\hat{U}$ and $\hat{V}$ are the vectors associated with functions u and v in S_n, we have

$$\hat{V}^* Q \hat{U} = \int_\Omega u\overline{v} \, dx. \tag{2.12}$$

where

$$\hat{V}^* = (\bar{V}_1, \cdots, \bar{V}_n). \tag{2.13}$$

The coercivity condition (2.5b) implies that

$$\mathrm{Re}\,(\hat{U}^* A\,\hat{U}) \geq K_0 \hat{U}^* Q\,\hat{U}. \tag{2.14}$$

Because Q is a positive definite matrix, (2.14) implies that the system (2.10c) has a unique solution $\hat{U}_n$.

We consider a direct iterative method for the computation of $\hat{U}_n$, and hence of u_n. We write

$$A = M - N, \tag{2.15}$$

where M is nonsingular and, in some sense, it is easy to solve problems of the form $M\,\hat{U} = \hat{G}$. Let a first guess $\hat{U}^{(0)}$ be chosen. Succeeding iterates are given by

$$M\,\hat{U}^{(\nu)} = N\,\hat{U}^{(\nu-1)} + \hat{F}. \tag{2.16}$$

It is well known that this procedure is convergent for any initial guess if and only if the spectral radius

$$\begin{aligned} \rho &:= \max\{|\lambda| : \lambda \text{ is an eigenvalue of } M^{-1}N\} \\ &= \max\{|\lambda| : \det(\lambda M - N) = 0\} \end{aligned} \tag{2.17}$$

of $M^{-1}N$ satisfies $\rho < 1$.

Our problem is the following. Imagine a sequence $\{S_n\}$ of subspaces and the corresponding matrix problems

$$A_n \hat{U}_n = \hat{F}_n. \tag{2.16}$$

where we have now used subscripts n on the problem matrices A_n and the moment vectors $\hat{F}_n$ to emphasize this sequence. Suppose the splittings $A_n = M_n - N_n$ of (2.15) are chosen in some regular fashion. We seek to determine the asymptotic behavior of the corresponding spectral radius ρ_n as $n \to \infty$.

3. The classical iterative methods.

Suppose A is an $n \times n$ matrix. The block structure of a direct iterative scheme for the problem

$$AX = Y \tag{3.1}$$

is completely determined by a block partition of the n-vector X. Suppose every vector X is decomposed into subvectors

$$X = (X_1, X_2, \cdots, X_r)^t$$

and each X_j is itself an n_j-vector. This partition of X induces a block partition $A = [A_{i,j}]$ in which each $A_{i,j}$ is an $n_i \times n_j$ matrix. The corresponding block Jacobi iterative scheme is

$$A_{i,i}X_i^{(\nu)} = -\sum_{s \neq i} A_{i,s}X_s^{(\nu-1)} + Y_i. \tag{3.2}$$

In terms of (2.16), M is the block diagonal matrix $M := \text{diag}[A_{i,i}]$. The corresponding Gauss-Seidel scheme is

$$A_{i,i}X_i^{(\nu)} = -\sum_{s<i} A_{i,s}X_s^{(\nu)} - \sum_{s>i} A_{i,s}X_s^{(\nu-1)} + Y_i, \tag{3.3}$$

while the SOR scheme with relaxation parameter ω is

$$A_{i,i}X_i^{(\nu)} = -\omega\sum_{s<i} A_{i,s}X_s^{(\nu)} - \omega\sum_{s>i} A_{i,s}X_s^{(\nu-1)} + \omega Y_i + (1-\omega)A_{i,i}X_i^{(\nu-1)}. \tag{3.4}$$

We will be interested in specific block structures that arise in a natural geometric way.

4. A general approach.

Our analysis of the iterative scheme (2.16) is an extension of the approach taken in [13] and [15]. We make four basic assumptions.

A.1 $\rho < 1$, so the iterative scheme is convergent.

A.2 ρ is an eigenvalue of $M^{-1}N$: there is a mesh vector $\hat{V} \neq 0$ such that

$$\rho M \hat{V} = N \hat{V}.$$

A.3 There are a constant q_0 and a function $q \in C^1(\bar{\Omega})$ with

$$q(x) \geq q_0 > 0 \quad (x \in \bar{\Omega}),$$

and a function $\eta(t)$, defined for $t \geq 1$ and satisfying

$$\eta(t) \to 0 \quad \text{as } t \to \infty,$$

such that for every u and v in S_n we have

$$h^{2m} \hat{V}^* N \hat{U} = \int q u \bar{v}\, dx + \varepsilon_n(u,v),$$

where

$$|\varepsilon_n(u,v)| \leq \eta(n)[\|u\|_1^2 + \|v\|_1^2 + \|u\|_1 + \|v\|_1].$$

A.4 Let q be the function of A.3. The eigenvalue problem that seeks $\lambda \in \mathbf{C}$ and $\varphi \in \tilde{H}^m$ to satisfy

$$B(\varphi,v) = \lambda \int q \varphi \bar{v}\, dx \quad \text{for all } v \in \tilde{H}^m \tag{4.1}$$

has a minimal eigenvalue

$$\Lambda_m = \Lambda_0 + iT.$$

By *minimal* we mean that for any eigenvalue λ it is true that $0 < \Lambda_0 \le \mathrm{Re}\,\lambda$, and that if $\mathrm{Re}\,\lambda = \Lambda_0$ then $|\lambda| \ge |\Lambda_m|$.

Note that if Λ_m is a minimal eigenvalue then so is $\overline{\Lambda_m}$; if $T = 0$, then

$$\Lambda_m = \Lambda_0 \le |\lambda| \quad \text{for any eigenvalue } \lambda.$$

Observe also that the eigenvalue problem (4.1) is equivalent to the problem

$$L\varphi = \lambda q \varphi \quad \text{in } \Omega, \quad b_j \varphi = 0 \quad \text{on } \partial\Omega \quad (0 \le j \le m-1).$$

Condition A.4 actually asserts that there is at least one eigenvalue. In the self-adjoint case, and in the case of a second order operator with Dirichlet boundary conditions, A.4 is always valid and $\Lambda_m = \Lambda_0$. We surmise that A.4 is always true, but we prefer to make the assumption explicit. Conditions A.1 and A.2 are readily verified for self-adjoint problems for which the splitting (2.15) satisfies block property A; see e.g. [1], [13]. For standard finite difference approximations of general second order Dirichlet problems, A.1 and A.2 follow from the Perron-Frobenius theory of positive matrices: see [20] and [15].

While one should write M_n, N_n, ρ_n, and h_n, we will usually drop the subscript n when its use is not essential for the clarity of the discussion.

Let $\lambda \ne 0$ be an eigenvalue of the iterative scheme (2.16), and let $\hat{W} \ne 0$ be an associated eigenvector, so that

$$\lambda M \hat{W} = N \hat{W}. \tag{4.2}$$

Subtract $\lambda N \hat{W}$ from both sides and divide by λ to see that

$$A\hat{W} = \frac{1-\lambda}{\lambda} N\hat{W}. \tag{4.3}$$

Now set

$$w := \sum \hat{W}_j \varphi_j, \quad \mu := \frac{1-\lambda}{\lambda h^{2m}}. \tag{4.4}$$

Then the eigenvalue problem (4.3) can be restated in the equivalent forms

$$\begin{aligned} &A\hat{W} = \mu(h^{2m}N)\hat{W}, \\ &B(w,v) = \mu \hat{V}^*(h^{2m}N)\hat{W} \quad \text{for all } v \in S_n. \end{aligned} \tag{4.5}$$

A basic result about this eigenvalue problem is

LEMMA 4.1. Suppose A.3-A.4 hold.

(a) Let $\{\mu_n\}$ be a bounded sequence of eigenvalues of (4.5), so that there is a constant $C > 0$ for which

$$|\mu_n| \le C.$$

Then the limit

$$\mu_\infty := \lim_{n' \to \infty} \mu_{n'}$$

of every convergent subsequence $\{\mu_{n'}\}$ is an eigenvalue of (4.1).

(b) Let Λ be an eigenvalue of (4.1), and fix $\delta > 0$. Then there is an n_1 so that for each $n \ge n_1$ there is an eigenvalue μ_n of (4.5) satisfying $|\Lambda - \mu_n| \le \delta$.

Proof. This result is essentially contained in the general theory for the

spectral approximation of compact operators (see [11]). However, for the sake of completeness, and to indicate an approach that applies in more general situations (see [14]), we give the proof in the appendix; (a) is Lemma A.1 and (b) is Theorem A.4.

THEOREM 4.2. Suppose A.3-A.4 hold. Then

$$\rho = \rho_n \geq 1 - \Lambda_0 h^{2m} + o(h^{2m}). \tag{4.6}$$

Proof. Lemma 4.1b implies that there is a sequence of eigenvalues of problems (4.1), which we denote by $\{\mu_n\}$, that converges to Λ_m. Thus Re $\mu_n \to \Lambda_0$ and

$$\text{Re}\,(1 + \mu_n h_n^{2m}) > 1,$$

whence

$$\lambda := \lambda(\mu_n) := \frac{1}{1 + \mu_n h_n^{2m}}$$

is a well defined eigenvalue of (4.2), $|\lambda| < 1$, and

$$|\lambda(\mu_n)| = 1 - \Lambda_0 h_n^{2m} + o(h_n^{2m}).$$

Therefore the theorem follows from the definition of ρ.

THEOREM 4.3. Suppose A.1-A.4 hold. Then

$$\rho = \rho_n = 1 - \Lambda_0 h_n^{2m} + o(h_n^{2m}). \tag{4.7}$$

Proof. Set

$$\tilde{\mu} := \frac{1 - \rho}{\rho h^{2m}}. \tag{4.8}$$

From A.1 and Theorem 4.2 we then see that

$$0 < \tilde{\mu} \leq \Lambda_0 + o(1). \tag{4.9}$$

More important, A.2 implies that $\tilde{\mu}$ is an eigenvalue of (4.5). From Lemma 4.1 and the definition of Λ_m we then have

$$\Lambda_0 + o(1) = \operatorname{Re} \tilde{\mu} = \tilde{\mu} \leq \Lambda_0 + o(1).$$

Therefore $\tilde{\mu} \to \Lambda_0$ as $n \to \infty$, and (4.7) holds.

REMARK. Because $\tilde{\mu}$ is an eigenvalue of (4.5), Lemma 4.1 shows that Λ_0 is itself an eigenvalue. Hence $\Lambda_m = \Lambda_0$ is real when A.1-A.4 hold.

Theorem 4.2 and A.2 suggest the following condition.

B.2 There is a constant C_0 for which, for every n, there is an eigenvalue λ_n that satisfies

$$|\lambda_n| = \rho_n \quad \text{and} \quad \left|\frac{1 - \lambda_n}{h_n^{2m}}\right| \leq C_0. \tag{4.10}$$

In fact, this condition can replace both A.1 and A.2.

THEOREM 4.4. Suppose B.2, A.3, and A.4 hold. Then the method is convergent and

$$\rho = 1 - \Lambda_0 h_n^{2m} + o(h_n^{2m}).$$

Proof. Let $\lambda_n = a_n + ib_n$. Then (4.10) implies

$$(1 - a_n)^2 + b_n^2 \leq h^{4m} C_0^2. \tag{4.11}$$

Set

$$\hat{\mu}_n := \frac{1 - \lambda_n}{\lambda_n h_n^{2m}}. \tag{4.12}$$

Then $\hat{\mu}_n$ is an eigenvalue of (4.5), and using (4.10) we have $|\hat{\mu}_n| \le 2C_0$ for h_n sufficiently small. Hence by Lemma 4.1 there is an eigenvalue Λ_1 of (4.1) and a subsequence $\{n'\}$ such that $\hat{\mu}_{n'} \to \Lambda_1$ as $n' \to \infty$. Let

$$\Lambda_1 = c + id. \tag{4.13}$$

Convergence of the subsequence and (4.12) together imply that

$$1 - \lambda_{n'} = \Lambda_1 \lambda_{n'} h_{n'}^{2m} + o(h_{n'}^{2m}).$$

Thus

$$\begin{aligned} \lambda_{n'} = a_{n'} + ib_{n'} &= 1 - \Lambda_1 \lambda_{n'} h_{n'}^{2m} + o(h_{n'}^{2m}) \\ &= 1 - (a_{n'} c - b_{n'} d) h_{n'}^{2m} - i(a_{n'} d + b_{n'} c) h_{n'}^{2m} + o(h_{n'}^{2m}). \end{aligned} \tag{4.14}$$

From (4.14) and the fact that

$$a_{n'} = 1 + O(h_{n'}^{2m}), \quad b_{n'} = O(h_{n'}^{2m})$$

we deduce that

$$a_{n'} = 1 - a_{n'} c h_{n'}^{2m} + o(h_{n'}^{2m}), \quad |\lambda_{n'}|^2 = 1 - 2c h_{n'}^{2m} + o(h_{n'}^{2m}).$$

Therefore

$$\rho_{n'} = |\lambda_{n'}| = 1 - c h_{n'}^{2m} + o(h_{n'}^{2m}).$$

By definition of Λ_c, $0 < \Lambda_0 \leq c = \mathrm{Re}\,\Lambda_1$, and therefore Theorem 4.2 implies that $c = \Lambda_c$. This proves the theorem.

5. A convergence theorem.

In earlier work [12], [13], and [15] for finite difference equations, an analogue of Theorem 4.3 established the asymptotic behavior of ρ. In this section we use Theorem 4.4 to obtain a new convergence theorem for finite element methods and splittings (2.15) that satisfy reasonable conditions. A remarkable feature of this proof is that we require no positivity, positive definiteness, or self-adjointness of the matrices involved. Therefore the theorem is particularly useful for nonself-adjoint problems and finite element discretizations. Moreover, the theorem can be recast to provide new results for finite difference approximations.

THEOREM 5.1. Consider the splitting (2.15) and the iterative scheme (2.16). Suppose A.3 holds and

$$|\varepsilon_n(u,u)| \le K_2[h\|u\|_0\|\nabla u\|_0 + h^2\|\nabla u\|_0^2], \tag{5.1}$$

$$|\mathrm{Im}\,(\hat{U}^* h^{2m} N \hat{U})| = |\hat{U}^* h^{2m} \frac{(N - N^*)}{2} \hat{U}| \le K_3 h(\|u\|_0^2 + |\varepsilon_n(u,u)|). \tag{5.2}$$

We also assume that certain "inverse inequalities" are satisfied: there are constants c_j such that

$$c_j h^j |u|_j \le |u|_0 \quad (j = 0, 1, 2, \cdots, m), \tag{5.3}$$

where $|u|_j$ is the seminorm of u defined by (1.10). Finally, we assume there is an eigenpair (λ,u) for which

$$|\lambda| = \rho, \quad \|u\|_m = 1, \quad \mathrm{Re}\,(\hat{U}^* N \hat{U}) \ge 0. \tag{5.4}$$

Then the iterative scheme (2.16) is convergent and

$$\rho = 1 - \Lambda_0 h^{2m} + o(h^{2m}). \tag{5.5}$$

REMARKS. If the inverse inequalities (5.3) hold, then Landau's inequality implies that there is a constant $\bar{c}_0$ such that

$$h^m \|u\|_m \le \bar{c}_0 \|u\|_0. \tag{5.6}$$

This inverse inequality is valid for many of the usual finite element spaces S_n: see [6]. Condition (5.4) holds when the splitting (2.15) is a block Jacobi splitting that satisfies block property A. This is so because block property A implies that $-\lambda$ is an eigenvalue of the iterative method whenever λ is: see [23, chapter 5]. The estimate (5.1) is a special form of the basic estimate of A.3. As we will see in section 7, precisely this estimate is satisfied in our model problem. Further, the derivation of this estimate and the arguments of [15, section 9] suggest that this is just the form to be expected. The estimate (5.2) arises because the antisymmetric part of N is usually related to the lower order terms of the elliptic operator L.

For the proof we need three lemmas.

LEMMA 5.2. Let $\vartheta_0 > 0$ be a fixed (small) constant. Then there is a constant C_{ϑ_0}, depending on ϑ_0 and m, and a constant $h(m)$, such that for every $u \in H^m(\Omega)$ and $h \le h(m)$ we have

$$h^2 \|\nabla u\|_0^2 \le \vartheta_0 \|u\|_0^2 + h^{2m} C_{\vartheta_0} |u|_m^2. \tag{5.7a}$$

$$h^2 \|\nabla u\|_0^2 \le \vartheta_0 \|u\|_0^2 + h^{2m} C_{\vartheta_0} \|u\|_m^2. \tag{5.7b}$$

Proof. We first derive a special form of Landau's inequality: for every $\vartheta > 0$ there are constants $D_\vartheta(k)$ and $\hat{h}$, depending only on ϑ and k, such that for every $u \in H^{k+1}$ and $h < \hat{h}$

$$|u|_k^2 \le \frac{\vartheta}{h^{2k}}\|u\|_0^2 + D_\vartheta(k)h^2|u|_{k+1}^2. \tag{5.8}$$

The proof of (5.8) follows by induction. The usual form of Landau's inequality (see [10]) is: there are constants c and α_0, depending only on Ω, so that for every positive α less than α_0

$$|u|_{j+1}^2 \le c(\alpha|u|_{j+2}^2 + \frac{1}{\alpha}|u|_j^2). \tag{5.9}$$

Hence (5.8) is true for $k = 1$. Assume (5.8) holds for $k = 1, 2, 3, \cdots, j$. Let

$$\alpha_1 := c \max\{2, 2D_\vartheta(k) : 1 \le k \le j\} \tag{5.10a}$$

and set

$$\alpha := \alpha_1 h^2. \tag{5.10b}$$

For some positive $\hat{h}_j$ it is true that $\alpha \le \alpha_0$ whenever $h < \hat{h}_j$. Then (5.9) yields

$$|u|_{j+1}^2 \le c(\alpha_1 h^2|u|_{j+2}^2 + \frac{1}{\alpha_1 h^2}|u|_j^2).$$

Using the inductive assumption we have

$$|u|_{j+1}^2 \le c(\alpha_1 h^2|u|_{j+2}^2 + \frac{1}{\alpha_1 h^2}[\frac{\vartheta}{h^{2j}}\|u\|_0^2 + D_\vartheta(j)h^2|u|_{j+1}^2]).$$

That is,

$$|u|_{j+1}^2 \le c\alpha_1 h^2 |u|_{j+2}^2 + \frac{c\vartheta}{\alpha_1 h^{2j+2}} \|u\|_0^2 + \frac{cD_\vartheta(j)}{\alpha_1} |u|_{j+1}^2.$$

But $cD_\vartheta(j)/\alpha_1 \le 1/2$ and $c/\alpha_1 \le 1/2$ by (5.10a), and so

$$|u|_{j+1}^2 \le 2c\alpha_1 h^2 |u|_{j+2}^2 + \frac{\vartheta}{h^{2(j+1)}} \|u\|_0^2.$$

Hence we have established (5.8) with $D_\vartheta(j+1) = 2c\alpha_1$. For $m = 1$ the inequality (5.7) is trivially true. For $m = 2$, the inequality (5.7) follows from (5.8) with $k = 1$. We proceed by induction. Suppose (5.7) is true for $m = 1, 2, \cdots, k$. Then for small positive $\bar{\vartheta}$ and for $h < \hat{h}_j(\bar{\vartheta})$ we have

$$h^2 |u|_1^2 \le C_{\bar{\vartheta}} h^{2k} |u|_k^2 + \bar{\vartheta} \|u\|_0^2. \tag{5.11}$$

Let

$$\bar{\vartheta} := \frac{\vartheta}{2}, \quad \vartheta_1 := \frac{\vartheta}{2C_{\bar{\vartheta}}(k)} = \frac{\bar{\vartheta}}{C_{\bar{\vartheta}}(k)}; \tag{5.12}$$

then (5.11) and (5.8) with $\vartheta = \vartheta_1$ together yield

$$h^2 |u|_1^2 \le C_{\bar{\vartheta}}(k) h^{2k} \left(D_{\vartheta_1} h^2 |u|_{k+1}^2 + \frac{\vartheta_1}{h^{2k}} \|u\|_0^2\right) + \bar{\vartheta} \|u\|_0^2$$

$$= C_{\bar{\vartheta}}(k) D_{\vartheta_1} h^{2(k+1)} |u|_{k+1}^2 + \frac{\vartheta}{2} \|u\|_0^2 + \frac{\vartheta}{2} \|u\|_0^2.$$

Setting $C_{\bar{\vartheta}}(k+1) := C_{\bar{\vartheta}}(k) D_{\vartheta_1}(k)$ completes the proof of the lemma.

COROLLARY. Assume that (5.1), (5.2), and (5.3) hold. Then there is a constant K_4 so that for small h and any $u \in S_n$ we have

$$|\mathrm{Im}\,(\hat{U}^* h^{2m} N \hat{U})| \le K_4 h \|u\|_0^2. \tag{5.13}$$

Furthermore, for any nonzero $u \in S_n$ set

$$a_0 := a_0(u) := \frac{|\varepsilon_n(u,u)|}{\|u\|_0^2}, \tag{5.14a}$$

$$\vartheta := \frac{a_0}{8K_2}, \quad \vartheta_0 := \min\{\vartheta, \vartheta^2\}. \tag{5.14b}$$

Assume that $a_0 \neq 0$ and set

$$k(a_0) := \frac{2C_{\vartheta_0}}{a_0}\left(1+\frac{1}{\vartheta}\right). \tag{5.14c}$$

Then for small enough h (5.3) implies that

$$c_m^2 h^{2m} |u|_m^2 \le \|u\|_0^2 \le k(a_0) h^{2m} \|u\|_m^2. \tag{5.15}$$

Thus if $u_n \in S_n$ satisfies

$$\|u_n\|_m = 1 \quad \text{and} \quad \frac{\|u_n\|_0^2}{h_n^{2m}} \to \infty, \tag{5.16}$$

then

$$a_0(u_n) \to 0 \quad \text{as } n \to \infty. \tag{5.17}$$

Proof. Estimate (5.13) follows from (5.1), (5.2), (5.3), and (5.7a). The lower bound of (5.15) is a restatement of (5.3). From (5.1) and (5.14a) we have

$$a_0\|u\|_0^2 \le K_2[\vartheta\|u\|_0^2 + (1 + \frac{1}{\vartheta})h^2|u|_1^2].$$

Using Lemma 5.2 we now deduce that

$$a_0\|u\|_0^2 \le K_2[(\vartheta + \vartheta_0 + \frac{\vartheta_0}{\vartheta})\|u\|_0^2 + C_{\vartheta_0}(1 + \frac{1}{\vartheta})h^{2m}\|u\|_m^2].$$

Our choice of ϑ and ϑ_0 implies that

$$K_2(\vartheta + \vartheta_0 + \frac{\vartheta_0}{\vartheta}) \le a_0/2.$$

Thus $\|u\|_0^2 \le k(a_0)h^{2m}\|u\|_m^2$, whence (5.15) and (5.17) follow.

For the rest of this section we assume that (5.1), (5.2), and (5.3) hold.

LEMMA 5.3. Let (λ,u) be the eigenpair of (5.4). Then there is a constant K_5 such that

$$|\hat{U}^* N \hat{U}| \ge K_5. \tag{5.18}$$

Proof. From $\lambda M\hat{U} = N\hat{U}$ it follows that

$$\lambda = \frac{\hat{U}^* N \hat{U}}{\hat{U}^* M \hat{U}} = \frac{\hat{U}^* N \hat{U}}{\hat{U}^* A \hat{U} + \hat{U}^* N \hat{U}}. \tag{5.19}$$

Evidently $|\hat{U}^* N \hat{U}| \ne 0$, because $\rho \ne 0$. Furthermore, the denominator is not zero because

$$\text{Re}\,(\hat{U}^* A \hat{U}) = \text{Re}\, B(u,u) \ge K_0\|u\|_m^2 = K_0 > 0$$

and u has been chosen so that $\text{Re}\,(\hat{U}^* N \hat{U}) \ge 0$. From (5.19) it follows that

$$\rho = |\lambda| = \frac{1}{|1 + \dfrac{\hat{U}^* A \hat{U}}{\hat{U}^* N \hat{U}}|}.$$

Now (5.18) must hold, for otherwise

$$|\frac{\hat{U}^* A \hat{U}}{\hat{U}^* N \hat{U}}| \to \infty,$$

which would violate Theorem 4.2.

Let (λ, u) be the eigenpair of (5.4). We write

$$\hat{U}^* h^{2m} N \hat{U} = \bar{q} + t_1 + i t_0, \tag{5.20a}$$

where

$$\bar{q} := \int q\,|u|^2 dx \tag{5.20b}$$

and t_1 and t_0 are real.

LEMMA 5.4. There are constants h_0, γ_1, and $\gamma_2 > 0$ so that for $0 < h < h_0$ we have

$$\gamma_2 \bar{q} \geq \bar{q} + t_1 = \mathrm{Re}\,(\hat{U}^* h^{2m} N \hat{U}) \geq \gamma_1 \bar{q} \geq \gamma_1 q_0 \|u\|_0^2. \tag{5.21}$$

Proof. Using (5.1) we obtain

$$|\varepsilon_n(u,u)| \leq K_2[\frac{1}{2}\|u\|_0^2 + \frac{3}{2}h^2|u|_1^2].$$

This inequality and (5.3), (5.4), and (5.7) together imply that there is a constant k' for which

$$|t_j| \le |\varepsilon_n(u,u)| \le K_2[\frac{1}{2}\|u\|_0^2 + \frac{3}{2}C\|u\|_0^2] =: k'\|u\|_0^2 \quad (j = 0, 1). \quad (5.22)$$

This establishes the upper bound of (5.21) with, say, $\gamma_2 := 1+k'/q_0$.

We now turn to the lower bound. If the lemma fails, then there is a subsequence $\{n'\}$ for which $(\lambda_{n'}, u_{n'})$ satisfies (5.4) and

$$\frac{h_{n'}^{-2m}(\bar{q}_{n'} + t_{1,n'})}{h_{n'}^{-2m}\bar{q}_{n'}} \to 0. \quad (5.23)$$

It follows from (5.6) that $\bar{c}_0^2 h^{-2m}\|u\|_0^2 \ge \|u\|_m^2 = 1$. Hence $h_{n'}^{-2m}\bar{q} \ge q_0/\bar{c}_0^2$, and so the numerator of (5.23), which is equal to $\mathrm{Re}\,(\hat{U}_{n'}^* N_{n'} \hat{U}_{n'})$, converges to 0. Lemma 5.3 then shows that $h_{n'}^{-2m}|t_0| = |\mathrm{Im}\,(\hat{U}_{n'}^* N_{n'} \hat{U}_{n'})| \ge K_5/2$. This last inequality reads

$$\frac{K_5}{2}h_{n'}^{2m} \le |t_0|. \quad (5.24)$$

We consider two cases. In the first, $h_{n'}^{-2m}\|u_{n'}\|_0^2 \to \infty$. But then (5.17) implies that

$$\frac{|t_{1,n'}|}{\bar{q}_{n'}} \le \frac{|t_{1,n'}|}{q_0\|u_{n'}\|_0^2} \to 0,$$

which contradicts (5.23). In the second, there is some constant C so that $\|u_{n'}\|_0^2 \le Ch_{n'}^{2m}$. But now (5.24) and (5.13) yield

$$\frac{K_5}{2}h_{n'}^{2m} \le |t_0| \le K_4 Ch_{n'}^{2m+1},$$

which is impossible.

Let us now compute

$$x + iy := \frac{\hat{U}^* A \hat{U}}{\hat{U}^* N \hat{U}^*}. \tag{5.25}$$

Let

$$\hat{U}^* A \hat{U} = \alpha(1 + i\sigma), \quad \alpha \geq K_0, \tag{5.26a}$$

where $|\sigma|$ is bounded because $B(u,v)$ is angle bounded.

Then

$$x = h^{2m} \alpha \frac{\bar{q} + t_1 + \sigma t_0}{(\bar{q} + t_1)^2 + t_0^2}, \tag{5.26b}$$

$$y = h^{2m} \alpha \frac{\sigma(\bar{q} + t_1) - t_0}{(\bar{q} + t_1)^2 + t_0^2}. \tag{5.26c}$$

Proof of the Theorem. Using (5.2) and (5.22), we deduce that

$$|t_0| \leq hK\bar{q}.$$

Hence

$$x \geq h^{2m} \alpha \frac{\gamma_1 \bar{q} - |\sigma| hK\bar{q}}{(\bar{q} + t_1)^2 + t_0^2} > 0. \tag{5.27}$$

Convergence follows from these inequalities and the fact that

$$\lambda = \frac{1}{1 + x + iy}.$$

Moreover, Theorem 4.2 implies that

$$1 \geq |\lambda| = \frac{1}{(1+x)^2 + y^2} \geq 1 - \Lambda_0 h^{2m} + o(h^{2m}).$$

Therefore there is a constant C so that

$$1 \geq 1 + 2x + x^2 + y^2 - Ch^{2m}.$$

Because x is positive and $2x + x^2 + y^2 \leq Ch^{2m}$, we get

$$0 < x < Ch^{2m}/2. \tag{5.28}$$

From (5.21), (5.26b), and (5.27) we see that (5.28) implies that there is a constant C for which $\bar{q} + t_1 \geq C$. Then (5.26c) shows that $|y| \leq Ch^{2m}$. Thus there is a constant C_0 so that

$$\frac{|1-\lambda|^2}{h^{4m}} = \frac{1}{h^{4m}} \frac{x^2 + y^2}{(1+x)^2 + y^2} \leq C_0.$$

But this means that B.2 holds; now Theorem 4.4 implies (5.5).

6. The model problem: description.

The basic ideas are clearest in this simple, but relatively rich, setting. Let Ω be the open unit square

$$\Omega := \{(x,y) \in \mathbf{R}^2 : 0 < x, y < 1\}.$$

Consider the Dirichlet problem

$$Lu = f \quad \text{for all } (x,y) \in \Omega, \tag{6.1}$$

$$u = 0 \quad \text{for all } (x,y) \in \partial\Omega. \tag{6.2}$$

Here L is the second order uniformly elliptic operator with smooth coefficients given by

$$Lu := -[(au_x)_x + (bu_x)_y + (bu_y)_x + (cu_y)_y] + d_1u_x + d_2u_y + d_0u \tag{6.3}$$

where

$$a(x,y) \geq a_0 > 0 \tag{6.4a}$$

$$b^2(x,y) - a(x,y)c(x,y) \leq -a_0 < 0 \tag{6.4b}$$

$$d_0(x,y) \geq 0. \tag{6.4c}$$

The inequalities (6.4) assert that L is a uniformly elliptic operator.

Set

$$b(u,v) := au_x\bar{v}_x + (bu_x\bar{v}_y + bu_y\bar{v}_x) + cu_y\bar{v}_y + (d_1u_x + d_2u_y + d_0u)\bar{v}, \quad (6.5a)$$

and

$$B(u,v) := \int\int_\Omega b(u,v)\,dx\,dy. \quad (6.5b)$$

As in section 2, we assume in addition to (6.4) that there is a constant $K_0 > 0$ such that for all $u \in H_0^1(\Omega)$ we have

$$\text{Re } B(u,u) \geq K_0\|u\|_1^2. \quad (6.6)$$

Under these circumstances, the boundary value problem (6.1), (6.2) is equivalent to the weak form that seeks $u \in H_0^1(\Omega)$ for which

$$B(u,v) = \int\int_\Omega f(x,y)\overline{v(x,y)}\,dx\,dy =: F(v) \quad \text{for all } v \in H_0^1(\Omega). \quad (6.7)$$

We now take up the finite element solution of this problem. Let $P_x \geq 2$ and $P_y \geq 2$ be integers, and set

$$\Delta x := \frac{1}{P_x + 1}, \quad \Delta y := \frac{1}{P_y + 1}, \quad r := \frac{\Delta y}{\Delta x}. \quad (6.8)$$

For any function $G(x,y)$ defined on Ω we write

$$G_{i,j} := G(i\Delta x, j\Delta y). \quad (6.9)$$

The finite element space S_n is the space of ***tensor products of Hermite cubic splines*** based on this grid. Let

$$x_i := i\Delta x \quad (i = 0, 1, 2, \cdots, P_x+1), \quad (6.10a)$$

$$y_j := j\Delta y \quad (j = 0, 1, 2, \cdots, P_y+1). \tag{6.10b}$$

For each pair (i,j), $0 \le i \le P_x$, $0 \le j \le P_y$, define

$$e_{i,j} := \{(x,y) : x_i \le x \le x_{i+1}, y_j \le y \le y_{j+1}\}. \tag{6.11}$$

Hence the corners of the rectangle $e_{i,j}$ are the points (x_i,y_j) -- the lower left corner -- (x_{i+1},y_j), (x_{i+1},y_{j+1}), and (x_i,y_{j+1}), as in Figure 1.

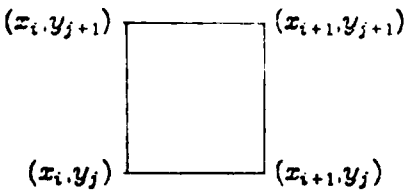


Figure 1. The element $e_{i,j}$.

The restriction of any function $v \in S_n$ to $e_{i,j}$ is a polynomial of degree three in each variable x and y, given by

$$v(x,y) = \sum_{\sigma,\mu=0}^{3} V_{\sigma,\mu} x^\sigma y^\mu \quad \text{for all } (x,y) \in e_{i,j}. \tag{6.12}$$

Thus v is determined in $e_{i,j}$ by the 16 parameters

$$\{v_{i+l,j+\rho}\}, \quad \{(v_x)_{i+l,j+\rho}\}, \quad \{(v_y)_{i+l,j+\rho}\}, \quad \{(v_{xy})_{i+l,j+\rho}\}, \tag{6.13}$$

with l and ρ running over the set $\{0, 1\}$.

Because we have been discussing the restriction of v to $e_{i,j}$ it might seem that we should somehow indicate that we are talking about v, v_x, v_y, v_{xy} *computed from within* $e_{i,j}$. However, the basic constraint on our space S_n is precisely that these values *at the four nodal points* are continuous. Therefore,

these four values can be associated with the geometric point (x_i, y_j).

It is convenient to describe S_n in terms of a local basis for the restriction to $e_{i,j}$. On the interval $0 \le x \le 1$ define the functions

$$V_0(x) := (1-x)^2(1+2x), \quad T_0(x) := x(1-x)^2,$$
$$V_1(x) := x^2(3-2x), \quad T_1(x) := (x-1)x^2. \tag{6.14}$$

These cubic polynomials satisfy

$$V_0(0) = 1, \quad V_0(1) = V'_0(0) = V'_0(1) = 0,$$
$$T'_0(0) = 1, \quad T_0(1) = T_0(0) = T'_0(1) = 0,$$
$$V_1(1) = 1, \quad V_1(0) = V'_1(0) = V'_1(1) = 0, \tag{6.15}$$
$$T'_1(1) = 1, \quad T_1(0) = T'_1(0) = T_1(1) = 0.$$

Then the restriction of v to $e_{i,j}$ may be written as

$$\begin{aligned} v(x,y) = &\sum_{l,\rho=0,1} v_{i+l,j+\rho} V_l\left[\frac{x-x_i}{\Delta x}\right] V_\rho\left[\frac{y-y_j}{\Delta y}\right] \\ &+ \sum_{l,\rho=0,1} \Delta x\,(v_x)_{i+l,j+\rho} T_l\left[\frac{x-x_i}{\Delta x}\right] V_\rho\left[\frac{y-y_j}{\Delta y}\right] \\ &+ \sum_{l,\rho=0,1} \Delta y\,(v_y)_{i+l,j+\rho} V_l\left[\frac{x-x_i}{\Delta x}\right] T_\rho\left[\frac{y-y_j}{\Delta y}\right] \\ &+ \sum_{l,\rho=0,1} \Delta x\,\Delta y\,(v_{xy})_{i+l,j+\rho} T_l\left[\frac{x-x_i}{\Delta x}\right] T_\rho\left[\frac{y-y_j}{\Delta y}\right]. \end{aligned} \tag{6.16}$$

Once we have this representation for the restriction of v to $e_{i,j}$ we can compute the "local mass matrix $Q(i,j)$" and the "local problem matrix $A(i,j)$." However, while we will compute some of these coefficients later, for our present purposes it suffices to observe the following form of these matrices. Let

$$V_{i,j} := (v_{i,j},\ \Delta x (v_x)_{i,j},\ \Delta y (v_y)_{i,j},\ \Delta x \Delta y (v_{xy})_{i,j})^t \tag{6.17}$$

be the 4-vector of interpolation conditions at the point (x_i, y_j), and let

$$\tilde{V}_{i,j} := (V^t_{i,j},\ V^t_{i+1,j},\ V^t_{i+1,j+1},\ V^t_{i,j+1})^t \tag{6.18}$$

be the 16-vector of all the interpolation conditions for $v(x,y)$ restricted to $e_{i,j}$. Then, for u and $v \in S_n$, with v represented by $\tilde{V}_{i,j}$ and u by $\tilde{U}_{i,j}$ on $e_{i,j}$, we have

$$\int\int_{e_{i,j}} u\bar{v}\, dx\, dy = \Delta x \Delta y\, \tilde{V}^*_{i,j} Q_0 \tilde{U}_{i,j}, \tag{6.19}$$

where Q_0 is a constant 16×16 matrix independent of Δx, Δy, or (i,j). Similarly,

$$\int\int_{e_{i,j}} b(u,v)\, dx\, dy = (\Delta x \Delta y)\tilde{V}^*_{i,j}[(\Delta x)^{-2}a_2 + (\Delta x)^{-1}a_1 + a_0]\tilde{U}_{i,j}, \tag{6.20}$$

where a_0, a_1, and a_2 are 16×16 matrices that depend on (i,j) and

$$a_0 \text{ is independent of } (\Delta x, \Delta y) \tag{6.21}$$

while

$$a_1 \text{ and } a_2 \text{ depend only on } r := \frac{\Delta y}{\Delta x}. \tag{6.22}$$

The matrix a_2 corresponds to the portion of $b(u,v)$ given by

$$a u_x \bar{v}_x + b(u_x \bar{v}_y + u_y \bar{v}_x) + c u_y \bar{v}_y \tag{6.23a}$$

and is a positive definite real symmetric matrix. The matrix a_1 corresponds to the portion of $b(u,v)$ given by

$$(d_1 u_x + d_2 u_y)\bar{v}, \tag{6.23b}$$

while the matrix a_0 corresponds to the portion of $b(u,v)$ given by

$$d_0 u \bar{v} \tag{6.23c}$$

and is a real symmetric positive semi-definite matrix.

From these local matrices one can easily construct the mass matrix Q and the problem matrix A; see for instance [19].

Let V_l denote the vector of all unknowns associated with the l'th horizontal line $y = y_l$: using (6.17),

$$V_l = (V^t_{1,l},\ V^t_{2,l},\ \cdots,\ V^t_{P_x,l})^t. \tag{6.24}$$

Now let $\hat{V}$ denote the vector of all unknowns ordered by lines, i.e.,

$$\hat{V} = (V^t_1,\ V^t_2,\ \cdots,\ V^t_{P_y})^t. \tag{6.25}$$

Then for any u and $v \in S_n$ we have

$$\int\int_\Omega u\bar{v}\, dx\, dy = \Delta x \Delta y\, \hat{V}^* Q \hat{U}, \tag{6.26}$$

where Q is a $(4P_xP_y)\times(4P_xP_y)$ constant matrix that is independent of Δx and Δy. Further, the finite element approximation (2.10c) corresponding to (2.7) with this choice of S_n and these interpolation conditions becomes

$$A\hat{U} = \hat{F}, \tag{6.27}$$

where

$$A = \Delta x \Delta y [(\Delta x)^{-2} A_2 + (\Delta x)^{-1} A_1 + A_0] \tag{6.28}$$

and A_0, A_1, and A_2 have the same qualitative features as a_0, a_1, and a_2. For example, A_2 corresponds to the portion of $b(u,v)$ given by (6.23a) and is a real symmetric positive definite matrix that depends on Δx and Δy only through the ratio $r = \Delta y / \Delta x$.

The matrices Q and A may be regarded as $(P_x P_y)\times(P_x P_y)$ block matrices, where each block is itself a 4×4 matrix. This is the "geometric point" representation of Q and A. In this representation both Q and A correspond to nine-point schemes. That is, the (i,j) block equations are

$$(A\hat{U})_{i,j} = A_{i,j;i,j} U_{i,j} + \sum_{l,\rho=-1,0,1} A_{i,j;i+l,j+\rho} U_{i+l,j+\rho} = F_{i,j}, \tag{6.29}$$

$$(Q\hat{U})_{i,j} = Q_{i,j;i,j} U_{i,j} + \sum_{l,\rho=-1,0,1} Q_{i,j;i+l,j+\rho} U_{i+l,j+\rho}. \tag{6.30}$$

We also consider the k-line representation of this problem. Let $k \geq 1$ be a fixed integer. We assume that k divides P_y, i.e.,

$$P_y = kP_k \tag{6.31}$$

where P_k is of course an integer. Let $u \in S_n$ and let $\hat{U}$ be the associated vector of interpolation conditions. Let $U_s(k)$ be the vector associated with the s'th set of k horizontal lines:

$$U_s(k) = (U_{k^2(s-1)+1}^t,\ U_{k^2(s-1)+2}^t,\ \ldots,\ U_{k^2s}^t)^t \tag{6.32}$$

In this representation, the problem matrix A and the mass matrix Q become block tridiagonal matrices. That is, the equation (2.16) takes the form

$$\begin{aligned} A_{s,s-1}(k)\,U_{s-1}(k) + A_{s,s}(k)\,U_s(k) + A_{s,s+1}(k)\,U_{s+1}(k) \\ = F_s(k) \quad (1 \le s \le P_k). \end{aligned} \tag{6.33}$$

The matrices $A_{s,s+\mu}(k)$ are $(4kP_s)\times(4kP_s)$ matrices.

Let us consider the block Jacobi iterative scheme based on this k-line representation of A. Given a first guess $\widetilde{U}^{(0)}$ we have the iterative scheme

$$A_{s,s-1}(k)\,U_{s-1}^{(\nu-1)}(k) + A_{s,s}(k)\,U_s^{(\nu)}(k) + A_{s,s+1}(k)\,U_{s+1}^{(\nu-1)} = F_s(k). \tag{6.34}$$

Thus, in the notation of section 2, we have (2.15) with

$$M := \mathrm{diag}[A_{s,s}(k)], \quad N := [-A_{s,s-1}(k),\ 0,\ -A_{s,s+1}(k)], \tag{6.35}$$

where the notation in (6.35) means that N is a block tridiagonal matrix with the three principal diagonals as written.

7. The model problem: estimates.

In this section and the next we turn to an analysis of block iterative methods for this model problem -- with a complete analysis of k-line block methods. Our first goal is to simplify the study of the bilinear form

$$\Delta x \Delta y \hat{V}^* N \hat{U}.$$

In particular, the estimates that follow enable us to ignore many of the elements of N when determining a function $q(x,y)$ that meets the conditions of A.3.

LEMMA 7.1. Let Ω be the unit square and let π be the 16-dimensional space of polynomials in $(x,y) \in \Omega$ that are cubic in each variable separately. Hence, if $g \in \pi$, then

$$g(x,y) = \sum_{r,s=0}^{3} g_{r,s} x^r y^s. \tag{7.1}$$

There is a constant $C_0 > 0$ such that

$$\frac{1}{4} \sum_{(\sigma,\mu)} |g(\sigma,\mu)|^2 \le C_0 \|g\|_0^2. \tag{7.2a}$$

$$\|g - g(\sigma,\mu)\|_0^2 \le C_0 \|\nabla g\|_0^2 = C_0 |g|_1^2. \tag{7.2b}$$

$$\sum_{(\sigma,\mu),(\sigma',\mu')} |g(\sigma,\mu) - g(\sigma',\mu')|^2 \le C_0 \|\nabla g\|_0^2. \tag{7.2c}$$

$$\sum_{|\alpha|=1}^{6} \sum_{(\sigma,\mu)} |(D^\alpha g)(\sigma,\mu)|^2 \le C_0 \|\nabla g\|_0^2. \tag{7.2d}$$

where (σ,μ) and (σ',μ') are any of the four corner points (0,0), (1,0), (1,1), (0,1).

Proof. Because π is a finite-dimensional space, any seminorm $|\cdot|$ is dominated by any norm $\|\cdot\|$: there is a constant C so that $|g| \le C\|g\|$ for every $g \in \pi$. This establishes (7.2a). Fix a corner point (σ,μ). Consider the norm defined on π by

$$|||g|||^2_{(\sigma,\mu)} := |g(\sigma,\mu)|^2 + \|\nabla g\|^2_0. \tag{7.3}$$

Because π is a finite-dimensional space, there is a constant $C > 0$ such that

$$\|\tilde{g}\|^2_0 \le C|||\tilde{g}|||^2_{(\sigma,\mu)} \quad \text{for all } \tilde{g} \in \pi. \tag{7.4}$$

Set

$$\tilde{g}(x,y) := g(x,y) - g(\sigma,\mu). \tag{7.5}$$

Then (7.2b) follows from (7.4) with C in place of C_0.

Let (σ',μ') be another corner point, and set

$$S_{(\sigma',\mu')}(g) := |g(\sigma',\mu')|.$$

Then $S_{(\sigma',\mu')}$ is a seminorm and, as before,

$$S^2_{(\sigma',\mu')}(\tilde{g}) \le C_1|||\tilde{g}|||^2_{(\sigma,\mu)} \quad \text{for all } \tilde{g} \in \pi. \tag{7.6}$$

Let $\tilde{g}$ be given by (7.5). Then (7.6) yields

$$|g(\sigma',\mu') - g(\sigma,\mu)|^2 \le C_1\|\nabla g\|^2_0.$$

Summing this inequality over all pairs (σ,μ), (σ',μ'), we obtain (7.2c) with C_0

replaced by $16C_1$.

Let S be the seminorm defined by

$$S^2(g) := \sum_{|\alpha|=1}^{6} \sum_{\sigma',\mu'=0}^{1} |(D^\alpha g)(\sigma',\mu')|^2.$$

The argument given above now yields (7.2d). Let C_0 be the largest of the constants and the lemma is proven.

For convenience, we collect some definitions here. Throughout this section and the next, we set

$$h := \sqrt{\Delta x\, \Delta y}.$$

For each $e_{i,j}$ and for every u and v in S_n, let

$$\eta(u,v,h,i,j) := h(\|u\|_{0;e_{i,j}}\|\nabla v\|_{0;e_{i,j}} + \|v\|_{0;e_{i,j}}\|\nabla u\|_{0;e_{i,j}}) + h^2\|\nabla u\|_{0;e_{i,j}}\|\nabla v\|_{0;e_{i,j}}.$$

Finally, for every u and v in S_n we set

$$\hat{\eta}(u,v,h) := h(\|u\|_0\|\nabla v\|_0 + \|v\|_0\|\nabla u\|_0) + h^2\|\nabla u\|_0\|\nabla v\|_0.$$

Note that

$$\sum_{e_{i,j}} \eta(u,v,h,i,j) \le \hat{\eta}(u,v,h).$$

LEMMA 7.2. Fix (i,j) and let $v \in S_n$. Let P and P' denote one of the corner points (x_i,y_j), (x_{i+1},y_j), (x_{i+1},y_{j+1}), (x_i,y_{j+1}) of $e_{i,j}$. Then there is a constant C_1 depending on r and $1/r$ such that

$$h^2 \sum_P |v(P)|^2 \le C_1 \|v\|^2_{0;e_{i,j}}. \tag{7.7a}$$

$$\| v - v(P) \|^2_{0;\sigma_{i,j}} \leq h^2 C_1 \| \nabla v \|^2_{0;\sigma_{i,j}} = h^2 C_1 | v |^2_{1;\sigma_{i,j}}. \tag{7.7b}$$

$$h^2 \sum_{|\alpha|=1}^{6} \sum_{P} \Delta x^{2\alpha_1} \Delta y^{2\alpha_2} | (D^\alpha v)(P) |^2 \leq C_1 h^2 \| \nabla v \|^2_{0;\sigma_{i,j}}. \tag{7.8}$$

$$h^2 \sum_{P,P'} | v(P) - v(P') |^2 \leq C_1 h^2 \| \nabla v \|^2_{0;\sigma_{i,j}}. \tag{7.9}$$

Proof. Set

$$\xi := \frac{x - x_i}{\Delta x}, \quad \eta := \frac{y - y_j}{\Delta y}.$$

Let v be any function in S_n and set

$$g(\xi,\eta) := v(x_i + \xi \Delta x, y_j + \eta \Delta y) \quad (0 \leq \xi, \eta \leq 1).$$

The function $g(\xi,\eta)$ is an element of π. Moreover, a direct computation shows that

$$\int\int_{\sigma_{i,j}} | v_x |^2 dx\, dy = \frac{\Delta y}{\Delta x} \int\int_{\Omega} | g_\xi |^2 d\xi\, d\eta = r \int\int_{\Omega} | g_\xi |^2 d\xi\, d\eta,$$

$$\int\int_{\sigma_{i,j}} | v_y |^2 dx\, dy = \frac{\Delta x}{\Delta y} \int\int_{\Omega} | g_\eta |^2 d\xi\, d\eta = \frac{1}{r} \int\int_{\Omega} | g_\eta |^2 d\xi\, d\eta.$$

Let

$$\bar{r} := \max(r, 1/r).$$

Then

$$\| \nabla v \|^2_{0;\Omega_{i,j}} \leq \bar{r} \| \nabla g \|^2_0 \leq \bar{r}^2 \| \nabla v \|^2_{0;\Omega_{i,j}}. \tag{7.10}$$

Another computation shows that

$$\int\int_{\Omega_{i,j}} | v(x,y) |^2 \, dx \, dy = h^2 \int\int_{\Omega} | g(\xi,\eta) |^2 \, d\xi \, d\eta. \tag{7.11}$$

Therefore the inequality (7.2b) of Lemma 7.1 yields

$$\| v - v(P) \|^2_{0;\Omega_{i,j}} \leq C_0 \bar{r} h^2 \| \nabla v \|^2_{0;\Omega_{i,j}},$$

which proves (7.7b). Inequality (7.7a) follows from a similar change of variables and (7.2a).

If we define

$$D^{\alpha}_{x} := \frac{\partial^{|\alpha|}}{\partial x^{\alpha_1} \partial y^{\alpha_2}}, \quad D^{\alpha}_{\xi} := \frac{\partial^{|\alpha|}}{\partial \xi^{\alpha_1} \partial \eta^{\alpha_2}}, \tag{7.12a}$$

then we see that

$$\Delta x^{\alpha_1} \Delta y^{\alpha_2} D^{\alpha}_{x} v(x,y) = D^{\alpha}_{\xi} g(\xi,\eta). \tag{7.12b}$$

Thus (7.2d) of Lemma 7.1 yields (7.8) with $C_1 = C_0 \bar{r}$. Finally, (7.2c) of Lemma 7.1 yields (7.9).

COROLLARY 7.3. There is a constant C_2 independent of (i,j) such that

$$\frac{h^2}{4} (v^2_{ij} + v^2_{i+1,j} + v^2_{i+1,j+1} + v^2_{i,j+1}) = \| v \|^2_{0;\Omega_{i,j}} + \bar{\eta}(v, \Delta x) \tag{7.13}$$

where

$$| \bar{\eta}(v, \Delta x) | \leq C_2 \eta(v, v, h, i, j). \tag{7.14}$$

Proof. For each index $(i+l, j+\rho)$ we apply (7.7b) to obtain

$$\|v - v_{i+l,j+\rho}\|_{0;e_{i,j}} \le h\sqrt{C_1}\,\|\nabla v\|_{0;e_{i,j}}.$$

Hence the triangle inequality yields

$$|(\|v\|_{0;e_{i,j}} - h\,|v_{i+l,j+\rho}|)| \le h\sqrt{C_1}\,\|\nabla v\|_{0;e_{i,j}}$$

and

$$h\,|v_{i+l,j+\rho}| \le \|v\|_{0;e_{i,j}} + h\sqrt{C_1}\,\|\nabla v\|_{0;e_{i,j}}. \tag{7.15a}$$

Therefore

$$h^2|v_{i+l,j+\rho}|^2 = \|v\|^2_{0;e_{i,j}} + \tilde{E}(v,h),$$

where

$$|\tilde{E}(v,h)| \le C\eta(v,v,h,i,j).$$

We add the four equations for all $(i+l, j+\rho)$ and divide by 4 to obtain (7.13).

THEOREM 7.4. Let u and $v \in S_n$. Let $\varphi \in C^1(\bar{\Omega})$. Then there is a constant $K > 0$ such that

$$h^2 \sum_{i=1}^{P_x} \sum_{j=1}^{P_y} |v_{i,j}|^2 \le K\|v\|_0^2. \tag{7.16a}$$

$$\int\int_\Omega \varphi u\bar{v}\,dx\,dy = h^2 \sum \varphi_{i,j} u_{i,j} \bar{v}_{i,j} + \delta(u,v,\varphi), \tag{7.16b}$$

where

$$|\delta(u,v,\varphi)| \le K(1 + \|\nabla\varphi\|_\infty)\hat{\eta}(u+v,u-v,h). \tag{7.16c}$$

Furthermore,

$$h^2\sum_{i,j}\sum_{|\alpha|=1}^{6}|\Delta x^{\alpha_1}\Delta y^{\alpha_2}(D^\alpha u)_{i,j}|^2 \le Kh^2\|\nabla u\|_0^2,$$

$$h^2\sum_{i,j}\sum_{|\alpha|+|\beta|\ge 1}|\Delta x^{\alpha_1+\beta_1}\Delta y^{\alpha_2+\beta_2}(D^\alpha u)_{i,j}(D^\beta v)_{i,j}| \le K\eta(u,v,h). \tag{7.17}$$

Proof. To get (7.16a), sum (7.7a) over all elements $e_{i,j}$. The estimates (7.16b) and (7.16c) follow from (7.13) and (7.14) of Corollary 7.3 applied to $(u \pm v)$, together with simple estimates on

$$\int\int_{e_{i,j}} \varphi u\bar{v}\,dx\,dy - \varphi_{i,j}\int\int_{e_{i,j}} u\bar{v}\,dx\,dy.$$

The first estimate of (7.17) comes from (7.8) of Lemma 7.2.

In order to complete the proof of (7.17) we need only consider the case $\alpha = 0$, $|\beta| \ge 1$. We have

$$J := h^2\sum_{i,j}|u_{i,j}(D^\beta v)_{i,j}|\Delta x^{\beta_1}\Delta y^{\beta_2}|$$

$$\le (h^2\sum(u_{i,j})^2)^{1/2}(h^2\sum_{|\beta|\ge 1}|(D^\beta v)_{i,j}\Delta x^{\beta_1}\Delta y^{\beta_2}|^2)^{1/2}.$$

A direct computation from (7.7a) and (7.17a) gives

$$J \le Kh\|u\|_0\|\nabla v\|_0.$$

whence (7.17b) follows.

We are now ready to apply these estimates to the study of

$$\Delta x \Delta y N = h^2 N$$

and the determination of the function $q(x,y)$.

THEOREM 7.5. For every vector $\hat{V} = (V_{i,j})$, let $\hat{V}^0$ be the vector with

$$\hat{V}^0 := (V^0_{i,j}), \quad V^0_{i,j} := \begin{pmatrix} v_{i,j} \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{7.18}$$

Suppose N^0 is a matrix such that for every u and $v \in S_n$ we have

$$h^2 \hat{V}^{0*} N^0 \hat{U}^0 = h^2 \hat{V}^{0*} N \hat{U}^0 + \vartheta(u,v,\Delta x), \tag{7.19a}$$

where

$$|\vartheta(u,v,\Delta x)| \le C\hat{\eta}(u,v,h). \tag{7.19b}$$

Then if N^0 satisfies A.3 so does N.

Proof. From the estimate (7.17b) we see that

$$h^2 \hat{V}^{0*} N \hat{U}^0 = h^2 \hat{V}^* N \hat{U} + O(\hat{\eta}(u,v,h)).$$

Hence when verifying A.3 it suffices to consider $\hat{V}^0$, $\hat{U}^0$, and matrices N^0 that satisfy (7.19).

This theorem shows that, to apply Theorem 5.1, we need consider only the vectors $\hat{V}^0$ and $\hat{U}^0$, and matrices N^0 satisfying (7.19). In particular, we may ignore components that come from derivative terms.

Suppose that the splitting (2.15) is a "natural" block splitting of the kind described in Section 3. That is, a nonzero element of $-N$ is exactly equal to the same (same indices (i,j)) nonzero element of A. Then it follows from Theorem 7.5 that *for the purpose of determining q and verifying* A.3, *it suffices to consider the same splitting of the matrix* A^0, which is a block matrix consisting of 4×4 blocks $A^0_{i,j;\sigma,\mu}$, where

$$A^0_{i,j;\sigma,\mu} = \Delta x \Delta y \begin{bmatrix} \alpha_{i,j;\sigma,\mu} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{7.20}$$

and

$$\alpha_{i,j;i,j} = \left[\frac{a_{i,j}}{\Delta x^2} + \frac{c_{i,j}}{\Delta y^2} \right] \frac{312}{175}, \tag{7.21a}$$

$$\alpha_{i,j;i+1,j} = \left[\frac{c_{i,j}}{\Delta y^2} \frac{54}{175} - \frac{a_{i,j}}{\Delta x^2} \frac{156}{175} \right], \tag{7.21b}$$

$$\alpha_{i,j;i+1,j+1} = -\left[\frac{b_{i,j}}{2\Delta x \Delta y} + \frac{27}{175} \left(\frac{a_{i,j}}{\Delta x^2} + \frac{c_{i,j}}{\Delta y^2} \right) \right], \tag{7.21c}$$

$$\alpha_{i,j;i,j+1} = \left[\frac{a_{i,j}}{\Delta x^2} \frac{54}{175} - \frac{c_{i,j}}{\Delta x^2} \frac{156}{175} \right], \tag{7.21d}$$

$$\alpha_{i,j;i-1,j+1} = -\left[\frac{27}{175} \left(\frac{a_{i,j}}{\Delta x^2} + \frac{c_{i,j}}{\Delta y^2} \right) \right] + \frac{b_{i,j}}{2\Delta x \Delta y}, \tag{7.22a}$$

$$\alpha_{i,j;i-1,j} = \alpha_{i,j;i+1,j}, \tag{7.22b}$$

$$\alpha_{i,j;i-1,j-1} = \alpha_{i,j;i+1,j+1}, \tag{7.22c}$$

$$\alpha_{i,j;i,j-1} = \alpha_{i,j;i,j+1}, \tag{7.22d}$$

$$\alpha_{i,j;i+1,j-1} = \alpha_{i,j;i+1,j+1}. \tag{7.22e}$$

REMARKS. As (7.21) and (7.22) show, on each element $e_{i,j}$ we have approximated the variable coefficients a, b, and c by constant coefficients $a_{i,j}$, $b_{i,j}$, $c_{i,j}$, respectively. One could "center" the coefficients $a_{i,j}$, $b_{i,j}$, and $c_{i,j}$ in these equations. However, for the purpose of determining q this increased accuracy is irrelevant. Only the coefficients $a(x,y)$, $b(x,y)$, and $c(x,y)$ enter into the formulae. This follows from the discussion in Section 6 describing a_2, a_1, and a_0, or equivalently A_2, A_1, and A_0 -- see (6.20)-(6.23) and (6.28).

8. The function $q(x,y)$: two iterative schemes.

In this section we use the results of Section 7, together with arguments already developed for finite difference equations (see [15]), to determine the function q that satisfies the conditions of A.3. We consider both the k-line Jacobi iterative scheme described in Section 6 and the point Gauss-Seidel method.

The finite element spaces S_n described by (6.11)-(6.13) satisfy the inverse inequalities (5.3). Moreover, because the k-line Jacobi scheme satisfies block property A, condition (5.4) holds. Hence to apply Theorem 5.1 we need only to confirm (5.2) and to show that A.3 holds with an estimate of the form (5.1). Now (5.2) follows from observing that the nonzero coefficients in $(N-N^*)$ come from coefficients in the problem matrix A that originate in the term

$$\Delta y A_1$$

of (6.28); but this term is of size $O(h)$. To finish the study of the k-line Jacobi method, we determine q in the following theorem. Of course, we will use the matrix N^0 that comes from A^0 defined by (7.20)-(7.22). N^0 acts only on the vectors $\hat{V}^0$ and $\hat{U}^0$. We define the norm

$$\|N^0\|_h := \sup \{(\sum_{i,j}(N^0\hat{U}^0)_{i,j}^2)^{1/2} : \sum_{i,j} | U_{i,j}^0 |^2 = 1\}.$$

THEOREM 8.1. Consider the k-line block Jacobi method described by equations (6.34). There is a function q that satisfies A.3 and

$$q(x,y) = \frac{12}{5}\frac{1}{kr}c(x,y). \tag{8.1a}$$

Therefore

$$\rho = 1 - \Lambda_0 h^2 + o(h^2), \tag{8.1b}$$

where Λ_0 is the minimal eigenvalue of (4.1). Hence from (1.6) and (1.7)

$$\rho_J(k) = 1 - \frac{5k}{12}\Gamma_0 h^2 + o(h^2), \quad \rho_{GS}(k) = 1 - \frac{5k}{6}\Gamma_0 h^2 + o(h^2),$$
$$\rho_b = 1 - 2(\frac{5k}{6}\Gamma_0)^{1/2} h + o(h). \tag{8.1c}$$

Proof. Following the development in Section 6 we see that it suffices to consider the matrix

$$N^0 = -\left[A^0_{s,s-1}(k),\ 0,\ A^0_{s,s+1}(k)\right] \tag{8.2}$$

where $A_{s,s\pm1}(k)$ are $(4kP_s)\times(4kP_s)$ matrices of the form

$$A^0_{s,s-1}(k) := \begin{bmatrix} 0 & A^0_{k(s-1)+1,k(s-1)}(1) \\ 0 & 0 \end{bmatrix}. \tag{8.3a}$$

$$A^0_{s,s+1}(k) := \begin{bmatrix} 0 & 0 \\ A^0_{ks,ks+1}(1) & 0 \end{bmatrix}. \tag{8.3b}$$

The matrices $A_{ks+1,ks}(1)$ and $A_{ks,ks+1}(1)$ are the $(4P_s)\times(4P_s)$ matrices that arise in the case $k = 1$. These matrices are themselves block tridiagonal matrices of the form

$$A^0_{j,j\pm1}(1) = [A^0_{i,j;i-1,j\pm1},\ A^0_{i,j;i,j\pm1},\ A^0_{i,j;i+1,j\pm1}] \tag{8.3c}$$

where the matrices $A^0_{i,j;\mu,\nu}$ are the 4×4 matrices given by (7.20)-(7.22c).

A direct computation now yields

$$\hat{V}^{0*}N^0\hat{U}^0 = -\sum_{s=1}^{P_k}[V^*_{ks}A^0_{ks,ks+1}(1)U_{ks+1}] - \sum_{s=1}^{P_k}[V^*_{ks+1}A^C_{ks+1,ks}(1)U_{ks}]. \qquad (8.4)$$

We rewrite (8.4) as

$$\hat{V}^{0*}N^0\hat{U}^0 = -\sum_{s=1}^{P_k}V^*_{ks}(A^0_{ks,ks+1} + A^0_{ks,ks-1})U_{ks}$$

$$-\sum_{s=1}^{P_k}V^*_{ks}A^0_{ks,ks+1}(U_{ks+1} - U_{ks}) - \sum_{s=1}^{P_k}V^*_{ks}(A^0_{ks+1,ks} - A^0_{ks,ks-1})U_{ks} \qquad (8.5)$$

$$+\sum_{s=1}^{P_k}(V^*_{ks} - V^*_{ks+1})A^0_{ks+1,ks}U_{ks}.$$

It follows that

$$\hat{V}^{0*}N^0\hat{U}^0 = -\sum_{s=1}^{P_k}V^*_{ks}(A^0_{ks,ks+1} + A^0_{ks,ks-1})U_{ks} + E(\hat{V}^0,\hat{U}^0), \qquad (8.6a)$$

where

$$|E(\hat{V}^0,\hat{U}^0)| \le \|N^0\|_h(\sum|v_{i,j}|^2)^{1/2}(\sum|u_{i,j+1} - u_{i,j}|^2)^{1/2}$$

$$+ \|N^0\|_h(\sum|u_{i,j}|^2)^{1/2}(\sum|v_{i,j+1} - v_{i,j}|^2)^{1/2} \qquad (8.6b)$$

$$+ h(\|\nabla b\|_\infty + \tau\|\nabla a\|_\infty + \frac{1}{\tau}\|\nabla c\|_\infty)(\sum|v_{i,j}|^2)^{1/2}(\sum|u_{i,j}|^2)^{1/2}.$$

Using the defining equations (7.20)-(7.22c) we see that

$$\hat{V}^{0*}N^0\hat{U}^0 = \sum_{s=1}^{P_k}\sum_{i=1}^{P_z}\frac{54}{175}(ra_{i,ks} + \frac{1}{\tau}c_{i,ks})\bar{v}_{i,ks}(u_{i-1,ks} + u_{i+1,ks})$$

$$+ \sum_{s=1}^{P_k}\sum_{i=1}^{P_z}(\frac{1}{\tau}\frac{312}{175}c_{i,ks} - \frac{108}{175}ra_{i,ks})\bar{v}_{i,ks}u_{i,ks} + E(\hat{V}^0,\hat{U}^0). \tag{8.7}$$

Because

$$u_{i-1,ks} + u_{i+1,ks} = 2u_{i,ks} + (u_{i-1,ks} - u_{i,ks}) + (u_{i+1,ks} - u_{i,ks}),$$

we see that

$$\hat{V}^{0*}N^0\hat{U}^0 = \sum_{s=1}^{P_k}\sum_{i=1}^{P_z}(\frac{12}{5}\frac{1}{\tau}c_{i,ks})\bar{v}_{i,ks}u_{i,ks} + E(\hat{V}^0,\hat{U}^0) + E_1(\hat{V}^0,\hat{U}^0), \tag{8.8}$$

where

$$|E_1(\hat{V}^0,\hat{U}^0)| \le \frac{108}{175}\|ra + \frac{1}{\tau}c\|_\infty(\sum|v_{i,j}|^2)^{1/2}(\sum|u_{i-1,j} - u_{i,j}|^2)^{1/2}.$$

The estimates of Theorem 7.4 now show that

$$h^2(E(\hat{V}^0,\hat{U}^0) + E_1(\hat{V}^0,\hat{U}^0)) = O(\hat{\eta}(u,v,h)),$$

whence from (8.8)

$$h^2\hat{V}^{0*}N^0\hat{U}^0 = h^2\sum_{s=1}^{P_k}\sum_{i=1}^{P_z}(\frac{12}{5}\frac{1}{\tau}c_{i,ks})\bar{v}_{i,ks}u_{i,ks} + O(\hat{\eta}(u,v,h)).$$

To complete the proof of Theorem 8.1 we employ an argument of [15, section 5]. Let $j = 1, 2, \cdots, k$. Then

$$v_{i,ks+j} = v_{i,ks} + G_{i,ks+j}(\hat{V}^0), \quad u_{i,ks+j} = u_{i,ks} + G_{i,ks+j}(\hat{U}^0), \tag{8.9}$$

where for any vector $\hat{W}^0$ we define

$$G_{i,ks+j}(\hat{W}^0) := \sum_{\mu=1}^{j}(w_{i,ks+\mu} - w_{i,ks+\mu-1}).$$

Observe that

$$|G_{i,ks+j}(\hat{W}^0)| \le \tilde{G}_{i,s}(\hat{W}^0) := k^{1/2}(\sum_{\mu=1}^{k}|w_{i,ks+\mu} - w_{i,ks+\mu-1}|^2)^{1/2}. \qquad (8.10)$$

Therefore

$$v_{i,ks+j}u_{i,ks+j} = v_{i,ks}u_{i,ks} + B_{i,ks+j}(\hat{V}^0, \hat{U}^0), \qquad (8.11)$$

where

$$|B_{i,ks+j}(\hat{V}^0, \hat{U}^0)| \le |v_{i,ks}|\tilde{G}_{i,s}(\hat{U}^0) + |u_{i,ks}|\tilde{G}_{i,s}(\hat{V}^0) + \tilde{G}_{i,s}(\hat{U}^0)\tilde{G}_{i,s}(\hat{V}^0).$$

Thus

$$c_{i,ks+j}\bar{v}_{i,ks+j}u_{i,ks+j} = c_{i,ks}\bar{v}_{i,ks}u_{i,ks} + D_{i,ks+j}(\hat{U}^0, \hat{V}^0), \qquad (8.12)$$

where

$$|D_{i,ks+j}(\hat{U}^0, \hat{V}^0)| \le \|c\|_\infty |B_{i,ks+j}(\hat{V}^0, \hat{U}^0)| + \Delta y \|\nabla c\|_\infty |v_{i,ks}u_{i,ks}|.$$

Sum equation (8.12) for $j = 1, 2, \cdots, k$ and divide by k. We obtain

$$\frac{1}{k}\sum_j c_{i,ks+j}\bar{v}_{i,ks+j}u_{i,ks+j} = c_{i,ks}\bar{v}_{i,ks}u_{i,ks} + H(\hat{V}^0, \hat{U}^0, c) \qquad (8.13a)$$

where

$$|H(\hat{V}^0,\hat{U}^0,c)| \le \|c\|_\infty(|v_{i,k_2}|\tilde{G}_{i,2}(\hat{U}^0) + |u_{i,k_2}|\tilde{G}_{i,2}(\hat{V}^0))$$

$$+ \|c\|_\infty\tilde{G}_{i,2}(\hat{U}^0)\tilde{G}_{i,2}(\hat{V}^0) + \Delta y\|\nabla c\|_\infty|v_{i,k_2}||u_{i,k_2}|. \tag{8.13b}$$

Returning to (8.8), we have

$$\Delta x\Delta y\,\hat{V}^{0*}N^0\hat{U}^0 = \Delta x\Delta y\sum_{j=1}^{P_1}\sum_{i=1}^{P_2}\frac{12}{5}\frac{1}{kr}c_{i,j}\bar{v}_{i,j}u_{i,j} + \varepsilon(\hat{V}^0,\hat{U}^0), \tag{8.14}$$

where the definitions of $E(\hat{V}^0,\hat{U}^0)$, $E_1(\hat{V}^0,\hat{U}^0)$, and $H(\hat{V}^0,\hat{U}^0,c)$ yield

$$|\varepsilon(\hat{V}^0,\hat{U}^0)| \le K(r)\hat{\eta}(u,v,h), \tag{8.15}$$

and the constant $K(r)$ depends on $(r + 1/r)$, all the coefficients, and all their gradient magnitudes $\|\nabla a\|_\infty$, $\|\nabla b\|_\infty$, and $\|\nabla c\|_\infty$. The theorem now follows from the estimates of Section 7.

We now turn to the case of the point Gauss-Seidel iterative method. First, let us clarify our terminology. In some sense there are two such methods. In one case we think of the geometric point (x_i,y_j) and associate with each such point a 4-vector $V_{i,j}$. In the course of this point Gauss-Seidel scheme we must invert a 4×4 matrix at each point. Alternatively, one may also consider the usual Gauss-Seidel method, in which one inverts the main diagonal; hence at each step we invert a scalar.

However, the estimates of Section 7 and the argument of Theorem 7.5 show that for finding q these methods are the same, in the sense that they both yield the same function q.

Unfortunately, for the point Gauss-Seidel method we cannot verify (5.4). Hence, although we can determine q in general, presently we can assert only

that the point Gauss-Seidel method converges for self-adjoint problems. Convergence in this instance follows from application of classical principles; see for example [20]. Even in this case we cannot prove that (5.4) holds, so we cannot establish the upper bound on ρ. But in general we do have the lower bound of Theorem 4.2. The next theorem summarizes these results.

THEOREM 8.2. Let the unknowns be ordered lexicographically and consider the corresponding Gauss-Seidel iterative scheme

$$A_{i,j;i,j}\,U_{i,j}^{(\nu)} = -\left(A_{i,j;i-1,j}\,U_{i-1,j}^{(\nu)} + A_{i,j;i-1,j-1}\,U_{i-1,j-1}^{(\nu)}\right)$$

$$-\left(A_{i,j;i,j-1}\,U_{i,j-1}^{(\nu)} + A_{i,j;i+1,j-1}\,U_{i+1,j-1}^{(\nu)}\right) \tag{8.16}$$

$$-\left(A_{i,j;i+1,j}\,U_{i+1,j}^{(\nu-1)} + A_{i,j;i-1,j+1}\,U_{i-1,j+1}^{(\nu-1)} + A_{i,j;i,j+1}\,U_{i,j+1}^{(\nu-1)} + A_{i,j;i+1,j+1}\,U_{i+1,j+1}^{(\nu-1)}\right) + F_{i,j}.$$

There is a function q that satisfies condition A.3 and

$$q(x,y) = \frac{156}{175}\left[ra(x,y) + \frac{1}{r}c(x,y)\right]; \tag{8.17}$$

moreover,

$$\rho \geq 1 - \Lambda_0 h^2 + o(h^2), \tag{8.18}$$

where Λ_0 is the minimal eigenvalue of (4.1). Of course, when L is self-adjoint the point Gauss-Seidel method is convergent.

Proof. Theorem 4.3 implies (8.18). Using A^0 rather than A we see that

$$\hat{V}^0 \cdot N^0 \hat{U}^0 = \sum_{i,j}\left(\frac{27}{175}\left[ra_{i,j} + \frac{1}{r}c_{i,j}\right] - \frac{b_{i,j}}{2}\right)v_{i,j}u_{i-1,j+1}$$

$$+ \sum_{i,j} \left(\frac{27}{175} \left[r a_{i,j} + \frac{1}{r} c_{i,j} \right] + \frac{b_{i,j}}{2} \right) \bar{v}_{i,j} u_{i+1,j+1}$$

$$+ \sum_{i,j} \left(- \frac{54}{175} r a_{i,j} + \frac{156}{175} \frac{1}{r} c_{i,j} \right) \bar{v}_{i,j} u_{i,j+1}$$

$$+ \sum_{i,j} \left(\frac{156}{175} r a_{i,j} - \frac{54}{175} \frac{1}{r} c_{i,j} \right) \bar{v}_{i,j} u_{i+1,j}.$$

The proof now follows from an argument similar to -- but much simpler than -- the argument in Theorem 8.1.

9. Comments.

As we have mentioned in the introduction, the basic theorems of section 4 are closely related to earlier finite difference results. While Theorems 4.4 and 5.1 are far-reaching extensions of our earlier work that are important for higher order and nonself-adjoint problems, even in the self-adjoint case the finite element equations present difficulties. One of the difficulties arises from having to deal with the many interpolation parameters, e.g., the 36 terms

$$\{u_{i+l,j+\rho}\},\quad \{(u_x)_{i+l,j+\rho}\},\quad \{(u_y)_{i+l,j+\rho}\},\quad \{(u_{xy})_{i+l,j+\rho}\},$$

with l and ρ running over the set $\{-1, 0, 1\}$, of the finite element equations at a point (x_i, y_j). It would be difficult to find q and verify A.3 if one had to deal with all these terms. However, Theorem 7.4 allows one to restrict attention to the nine function values $\{u_{i+l,j+\rho}\}$.

This observation leads one to ask, are the estimates of Theorem 7.4 specific only to these cases, or can the estimates be obtained for a general class of finite element spaces? Looking at section 7, we see that there are two points essential to the development of Theorem 7.4.

(1) The interpolation parameters consist of function values and derivatives at certain vertices of the element $e_{i,j}$.

(2) There are a fixed and finite number of finite dimensional spaces $\pi_1, \pi_2, \cdots, \pi_R$ (in our case, $R = 1$), and for every S_n it is true that to each element $e_{i,j}$ there corresponds a fixed π_t and a smooth mapping for which the ***restrictions*** of the functions of S_n to $e_{i,j}$ are the images of π_t under the mapping. See Lemma 7.2. Furthermore, in π_t one obtains estimates like those of Lemma 7.1.

Therefore it is quite clear that this approach to the simplification of $(\hat{P}^* N \hat{U})$

will apply to many finite element spaces $\hat{S}_n$. In particular, it applies to all the nodal finite element spaces (see [18], [19]), provided that there is some regularity of the elements e whose union is Ω -- say, provided that the diameters of neighboring elements vary slowly. Now we ask, given that such estimates hold and the analysis of $(\hat{V}^* N \hat{U})$ is reduced to a study of $(\hat{V}^{0*} N^0 \hat{U}^0)$, which corresponds to a related generalized finite difference equation, should one expect to find a q for general domains and general elliptic equations? The answer appears to be yes! For the finite difference case, this point is discussed in [15, section 9]. Of course, if one really wants to determine q and hence the asymptotic form of ρ, one must work out the details in any particular case.

Even when q is known, the eigenvalue Λ_0 is not readily available. Hence one might question the practical value of this theory. However, there are at least four important ways in which the theory is useful.

(1) There are cases -- model problems -- in which one can compute the eigenvalues. For these model problems it is then possible to compare different methods.

(2) In second order elliptic problems with nice boundary conditions, and in general self-adjoint elliptic problems, the smallest eigenvalue is monotone decreasing in q: that is, if $q_1(x) \leq q_2(x)$ for all $x \in \Omega$, then $\Lambda_0(q_1) \geq \Lambda_0(q_2)$. In these instances qualitative comparisons of different methods are possible.

(3) Consider the k-line methods. Here the basic blocks are monotone in k. This fact is reflected in (8.1), where q is inversely proportional to k, so that $\Lambda_0 = Ck\Gamma_0$ is directly proportional to k. Thus (8.1a) and (1.5) hold. Hence we can compare k-line methods for different values of k even when we do not know the exact value of Γ_0. It is easy to imagine a situation where one has such cases of monotone blocks. We should then be able to compare

methods without knowing the basic Γ_0.

(4) Consider the relationship between the k-line Gauss-Seidel method and the point Gauss-Seidel method(s), which is revealed by comparing Theorems 8.1 and 8.2. By computing the work per sweep, one easily sees that -- in the best of circumstances, where ρ is as small as possible and equality holds in (8.18) -- the k-line Gauss-Seidel methods are to be preferred to the point Gauss-Seidel methods. For instance, in the simplest case where $r = 1$ and $a = c = 1$, we see that

$$\rho_{GS}(k) \approx 1 - \frac{5}{3} k \pi^2 h^2, \quad \rho_{GS}(point) \approx 1 - \frac{175}{156} \pi^2 h^2.$$

We now turn to another aspect of these results. For second order problems ($m = 1$), the basic Jacobi and Gauss-Seidel methods -- but not the SOR method -- have spectral radii $\rho \approx 1 - Kh^2$. The exponent 2 arises from the fact that the elliptic equation is of second order; it has nothing to do with the dimension of Ω or the order of accuracy of the discretization method, which in the case of Hermite cubic splines is 4. Thus for reasonably desired error tolerances, the finite element h is large compared to the usual finite difference h, and the finite element ρ is correspondingly smaller.

Block splittings based on geometrically natural blocks have a property that is important for exploiting new computer architectures: the corresponding problems (2.16) are easy to set up on vector and multiprocessor machines, because M decomposes into independent submatrices. The coupling between subregions of Ω is isolated in N. Hence overhead associated with data transmission between processors is small. Moreover, decomposition of (1.1) into (2.16) is easily managed by hand. This is a telling factor where operating systems, compilers, and other supporting software are unlikely to make available to users all

the resources of multiprocessor machines in a simple way. Leaving aside this practical point, we note simply that the independent subproblems of (2.16) can be attacked simultaneously by independent processors. Hence block decompositions permit parallel computation even as they provide improved convergence rates.

Appendix: the eigenvalue problem.

The major purpose of this section is to prove Lemma 4.1. Let us clarify the notation and restate the basic hypothesis.

With every function

$$u(x) = \sum U_i \varphi_i(x) \in S_n \tag{a.1a}$$

we associate the vector of coefficients of u

$$\hat{U} = (U_1, U_2, \cdots, U_n)^t. \tag{a.1b}$$

There are three basic matrices A, Q, and N, which satisfy

$$\hat{V}^* A \hat{U} = B(u,v) \quad \text{for all } u \text{ and } v \in S_n, \tag{a.2a}$$

$$\hat{V}^* Q \hat{U} = \int u\bar{v}\, dx \quad \text{for all } u \text{ and } v \in S_n, \tag{a.2b}$$

$$\hat{V}^* (h^{2m} N) \hat{U} = \int qu\bar{v}\, dx + \varepsilon_n(u,v) \quad \text{for all } u \text{ and } v \in S_n. \tag{a.3b}$$

Here $q \in C^1(\Omega)$, $q(x) \geq q_0 > 0$ on Ω,

$$|\varepsilon_n(u,v)| \leq \eta(n)[(1+\|u\|_1)(1+\|v\|_1) + \|u\|_1^2 + \|v\|_1^2], \tag{a.3c}$$

and

$$\eta(n) \to 0 \quad \text{as } n \to \infty. \tag{a.3d}$$

Note that A.3 implies (a.3c).

We are concerned with the eigenvalue problem

$$A\hat{U} = \mu h^{2m} N \hat{U}, \quad \hat{U} \neq 0 \tag{a.4a}$$

and its relationship to the eigenvalue problem

$$B(u,v) = \Lambda \int qu\bar{v}\, dx \quad \text{for all } v \in \tilde{H}^m. \tag{a.4b}$$

This latter problem is completely equivalent to the problem

$$Lu = \Lambda q u \quad \text{in } \Omega, \quad b_j u = 0 \quad \text{on } \partial\Omega \quad (0 \leq j \leq m-1). \tag{a.5}$$

LEMMA A.1 (Lemma 4.1a). Let $\{\mu_n\}$ be a bounded sequence of eigenvalues of (a.4a), so that there is a constant $C > 0$ for which

$$|\mu_n| \leq C. \tag{a.6}$$

Then the limit

$$\mu_\infty := \lim_{n' \to \infty} \mu_{n'} \tag{a.7}$$

of every convergent subsequence $\{\mu_{n'}\}$ is an eigenvalue of (a.4).

Proof. Let $\hat{U}(n)$ be the eigenvector corresponding to μ_n, normalized so that

$$\|u(n)\|_0 = 1. \tag{a.8a}$$

Thus

$$\hat{U}^*(n) A_n \hat{U}(n) = \mu_n \hat{U}^*(n) (h_n^{2m} N_n) \hat{U}(n). \tag{a.8b}$$

Using (a.3b) and (a.3c), we get

$$\hat{U}^*(n)A_n\hat{U}(n) = \mu_n\int q\,|u|^2\,dx + \varepsilon_n(u,u) \tag{a.9}$$

and

$$|\varepsilon_n(u,u)| \le \eta(n)(2 + 4\|u(n)\|_1^2).$$

Therefore from (2.5) we have

$$K_0\|u(n)\|_m^2 \le \mathrm{Re}\,(\hat{U}^*(n)A_n\hat{U}(n)) \le C\|q\|_\infty + \eta(n)(2 + 4\|u\|_1^2).$$

Hence for n large enough

$$\|u(n)\|_m^2 \le \frac{2C}{K_0}\|q\|_\infty. \tag{a.10a}$$

Now choose a convergent subsequence $\{\mu_{n'}\}$ and let its limit be μ_∞. By (a.10a) there is a subsequence $\{n''\}$ of $\{n'\}$ and a function $\varphi \in \tilde{H}^m$ so that

$$u(n'') \to \varphi \text{ weakly in } H^m(\Omega). \tag{a.10}$$

If $v \in \tilde{H}^m$ and $v(n)$ is its $H^m(\Omega)$ projection onto S_n, then from (a.4a) we have

$$\hat{V}^*(n)A_n\hat{U}(n) = \mu_n\hat{V}^*(n)(h_n^{2m}N)\hat{U}(n).$$

Passage to the limit along $\{n''\}$ yields

$$B(\varphi,v) = \mu_\infty\int q\varphi\bar{v}\,dx. \tag{a.11}$$

Hence either $\varphi \equiv 0$, or φ is an eigenfunction of (a.4) with corresponding eigen-

value μ_m. But the normalization (a.8a) implies that $\|\varphi\|_0 = 1$, and the lemma is proven.

In preparation for the proof of Lemma 4.1b we develop some additional results. Consider the inhomogeneous problem

$$B(u,v) = \int qf\bar{v}\,dx \quad \text{for all } v \in \mathring{H}^m. \tag{a.12a}$$

For any $f \in L^2(\Omega)$, the solution u of (a.12a) is in $\mathring{H}^m$. Let $T:L^2 \to \mathring{H}^m$ denote the solution operator

$$Tf = u. \tag{a.12b}$$

Similarly, let $T_n:S_n \to S_n$ defined by

$$T_n f = u \tag{a.13a}$$

denote the solution operator for the discrete inhomogeneous problem

$$A\hat{U} = h^{2m} N\hat{F}. \tag{a.13b}$$

Observe that while (a.13b) is stated in terms of the vectors $\hat{U}$ and $\hat{F}$, the operator T_n maps the function f to the function u.

Our goal is a discussion of the relationship of the spectra of these operators. If Λ is an eigenvalue of (a.4b) then $1/\Lambda$ is an eigenvalue of T; similarly, if μ is an eigenvalue of (a.4a) then $1/\mu$ is an eigenvalue of T_n.

LEMMA A.2. Let Σ be a bounded subset of the resolvent set of T with

$$0 \notin \Sigma. \tag{a.14}$$

Let $z \in \Sigma$ and let $f \in S_n$. Consider the inhomogeneous problem

$$A\hat{U} - \frac{1}{z}h^{2m}N\hat{U} = Q\hat{F}. \tag{a.15}$$

There is an integer n_0 and a constant K depending on Σ, but not on n or f, such that for $n \geq n_0$ (a.15) has a unique solution $u \in S_n$ and

$$\|u\|_0 \leq K\|f\|_0. \tag{a.16}$$

(Note once more that we pose the problem in terms of $\hat{U}$ and $\hat{F}$ but consider the solution as a function $u \in S_n$.)

Proof. Because (a.15) is a linear problem and S_n is a finite-dimensional space, the lemma follows once we have established (a.16). Suppose (a.16) is false. Then there is a subsequence $\{n'\}$ for which the complex number $z_n \in \mathbf{C}$ and the functions $f(n)$ and $u(n) \in S_n$ that are related by (a.15) satisfy

$$z_n \to z_\infty \in \Sigma, \quad \|u(n)\|_0 = 1, \quad \|f(n)\|_0 \to 0. \tag{a.17}$$

However, from (a.15) we have

$$\hat{U}^*(n)A_n\hat{U}(n) - \frac{1}{z_n}\hat{U}^*(n)(h_n^{2m}N_n)\hat{U}(n) = \hat{U}^*(n)Q\hat{F}(n).$$

We rewrite this to get

$$B(u(n),u(n)) - \frac{1}{z_n}\int q\,|u(n)|^2\,dx = \int f\bar{u}\,dx + \frac{1}{z_n}\varepsilon_n(u,u). \tag{a.18}$$

Arguing as in the proof of Lemma A.1, we find for n' large enough that

$$\|u(n)\|_m^2 \leq \frac{2\|q\|_\infty}{K_0|z|_{\min}} + 2\|f(n)\|_0.$$

Therefore there is a subsequence $\{n''\}$ of $\{n'\}$ and a function $\varphi \in \tilde{H}^m$ so that

$$u(n'') \to \varphi \text{ weakly in } H^m(\Omega), \quad \|\varphi\|_0 = 1.$$

Moreover, (a.17) and (a.18) imply that for every $v \in \tilde{H}^m$ the function φ satisfies

$$B(\varphi, v) = \frac{1}{z_-} \int q\, \varphi \bar{v}\, dx.$$

Hence $1/z_-$ is an eigenvalue of problem (a.4) and z_- is an eigenvalue of T. But this is impossible.

LEMMA A.3. Fix $g \in H^m(\Omega)$ and let $g(n) \in S_n$ be the L^2 projection of g onto S_n. Let Σ be as above. For every $z \in \Sigma$, let $u(x;z)$ and $u(x;z,n)$ be the solutions of

$$B(u, v) - \frac{1}{z} \int q u \bar{v}\, dx = \int g \bar{v}\, dx \quad \text{for all } v \in \tilde{H}^m,$$

$$A\hat{U}(z,n) - \frac{1}{z} h^{2m} N \hat{U}(z,n) = Q\hat{G}(n). \tag{a.19}$$

Then

$$\|u(\cdot;z,n)\|_0 \le K_0 \|g\|_0, \tag{a.20}$$

$$\|u(\cdot;z) - u(\cdot;z,n)\|_0 \to 0 \quad \text{as } n \to \infty.$$

Proof. For each n we have

$$\hat{U}^* A \hat{U} - \frac{1}{z} \hat{U}^* (h^{2m} N) \hat{U} = \hat{U}^* Q \hat{G}(n).$$

This equation may be rewritten as

$$B(u,u) - \frac{1}{z}\int q\,|u|^2\,dx = \int g\bar{u}\,dx + \frac{1}{z}\varepsilon_n(u,u).$$

Because $\|u\|_0$ is uniformly bounded, the usual argument shows that $u(\cdot;z,n) \to u(\cdot;z)$ weakly in $H^m(\Omega)$; but then $u(\cdot;z,n) \to u(\cdot;z)$ strongly in $L^2(\Omega)$.

THEOREM A.4. Let $\sigma = 1/\Lambda$ be an eigenvalue of T. Let $\delta > 0$ be chosen so small that σ is the only eigenvalue of T inside the circle about σ of radius 2δ, and this circle lies entirely in the right half plane Re $z > 0$. Then there is an n_1 so that for each $n \geq n_1$ there is an eigenvalue σ_n of T_n satisfying

$$|\sigma - \sigma_n| < \delta. \tag{a.21a}$$

Consequently $\mu_n = 1/\sigma_n$ satisfies

$$|\Lambda - \mu_n| \leq \frac{\delta|\Lambda|^2}{1 - |\Lambda|\delta}. \tag{a.21b}$$

Proof. We consider the two projection operators

$$E := \frac{1}{2\pi i}\oint_\Gamma (z - T)^{-1}\,dz, \quad E_n := \frac{1}{2\pi i}\oint_\Gamma (z - T_n)^{-1}\,dz, \tag{a.22}$$

where Γ is the circle $\Gamma = \{z \in C : |z - \sigma| = \delta\}$. In order to prove the theorem we need only show there is an n_1 such that for all $n \geq n_1$

$$E_n \neq 0. \tag{a.23}$$

Let φ be the eigenfunction associated with σ. Then

$$\sigma L\varphi = q\varphi \quad \text{in } \Omega, \quad b_j\varphi = 0 \quad \text{on } \partial\Omega \quad (0 \leq j \leq m-1). \tag{a.24}$$

Equivalently,

$$\sigma B(\varphi,v) = \int q\,\varphi\bar{v}\,dx \quad \text{for all } v \in \tilde{H}^m. \tag{a.25}$$

Moreover,

$$E\varphi = \varphi. \tag{a.26}$$

The function

$$u(\cdot;z) := (z - T)^{-1}\varphi \tag{a.27a}$$

satisfies

$$Lu - \frac{1}{z}qu = \frac{1}{z}L\varphi = \frac{1}{z\sigma}q\varphi \quad \text{in } \Omega, \quad b_j u = 0 \quad \text{on } \partial\Omega. \tag{a.27b}$$

In other words, for every $v \in \tilde{H}^m$ we have

$$B(u,v) - \frac{1}{z}\int q(x)u(x;z)\overline{v(x)}\,dx = \frac{1}{z\sigma}\int q\,\varphi\bar{v}\,dx. \tag{a.28}$$

Let $g(\cdot;n) \in S_n$ be the L^2 projection of $q\varphi$ onto S_n. Let $w(\cdot;n)$ be the solution of

$$B(w,v) = \frac{1}{\sigma}\int q\,\varphi\bar{v}\,dx \quad \text{for all } v \in S_n. \tag{a.29}$$

Then $\hat{W}$ satisfies

$$A\hat{W} = \frac{1}{\sigma}Q\hat{G} \tag{a.30}$$

Now set

$$\varphi(\cdot;n) := E_n w(\cdot;n). \tag{a.31}$$

Our goal is to show that there is an n_1 such that $\varphi(\cdot;n) \neq 0$ whenever $n \geq n_1$. We have

$$\varphi(\cdot;n) = \frac{1}{2\pi i} \oint_\Gamma (z - T_n)^{-1} w(\cdot;n)\, dz. \tag{a.32}$$

Let

$$v(\cdot;z,n) := (z - T_n)^{-1} w(\cdot;n). \tag{a.33}$$

A straightforward calculation shows that $\hat{V}$ satisfies

$$A\hat{V} - \frac{1}{z}(h^{2m}N)\hat{V} = \frac{1}{z}A\hat{W} = \frac{1}{z\sigma}Q\hat{G}. \tag{a.34}$$

Comparing (a.28) and (a.34), we see by Lemma A.3 that

$$\|v(\cdot;z,n)\|_0 \leq K \tag{a.35a}$$

for some constant K, and

$$\|v(\cdot;z,n) - u(\cdot;z)\|_0 \to 0. \tag{a.35b}$$

By (a.22), (a.26), and (a.27a),

$$\varphi = \frac{1}{2\pi i} \oint_\Gamma u(\cdot;z)\, dz; \tag{a.36a}$$

moreover, by (a.31) and (a.32)

$$\varphi(\cdot;n) = \frac{1}{2\pi i} \oint_\Gamma v(\cdot;z,n)\, dz. \tag{a.36b}$$

The dominated convergence theorem and (a.31) now imply that

$$\|\varphi(\cdot;n) - \varphi(\cdot)\|_0 \to 0 \quad \text{as } n \to \infty.$$

But $\varphi \neq 0$; hence $\varphi(\cdot;n) \neq 0$ for large n, and the theorem is proven.

REMARK. Following the argument in [14], one can give a complete discussion of the relationship of the eigenvalue problem (a.4a) to the eigenvalue problem (a.4b). All that remains for completeness is to establish that the multiplicity of eigenvalues is preserved. However, because it is not relevant to our present purposes we omit it.

References

[1] R. J. ARMS, L. D. GATES, AND B. ZONDEK, *A method of block iteration*, J. Soc. Ind. Appl. Math., **4** (1956), pp. 220-229.

[2] K. AZIZ AND I. BABUSKA, eds., *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, 1973.

[3] D. L. BOLEY AND S. V. PARTER, *Block relaxation techniques for finite element elliptic equations: an example*, LASL report LA-7870-MS (1979).

[4] C. DEBOOR, ed., *Mathematical Aspects of Finite Elements in Partial Differential Equations*, Academic Press, New York, 1974.

[5] J. H. BRAMBLE AND A. H. SCHATZ, *Rayleigh-Ritz-Galerkin methods for Dirichlet's problem using subspaces without boundary conditions*, Comm. Pure Appl. Math., **23** (1970), pp. 653-675.

[6] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, The Netherlands, 1978.

[7] G. J. FIX AND K. LARSEN, *On the convergence of SOR iterations for finite element approximations to elliptic boundary value problems*, SIAM J. Numer. Anal., **8** (1971), pp. 536-547.

[8] S. P. FRANKEL, *Convergence rates of iterative treatments of partial differential equations*, Math. Tables Aids Comp., **4** (1950), pp. 65-76.

[9] A. GEORGE, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal., **10** (1973), pp. 345-363.

[10] R. R. KALLMAN AND G.-C. ROTA, *On the inequality* $\|f'\|^2 \leq 4\|f\|\cdot\|f''\|$, in *Inequalities II*, O. SHISHA, ed., Academic Press, New York, 1970, pp. 187-191.

[11] J. E. OSBORN, *Spectral approximation for compact operators*, Math. Comp., **29** (1975), pp. 712-725.

[12] S. V. PARTER, *Multi-line iterative methods for elliptic difference equations and fundamental frequencies*, Numer. Math., **3** (1961), pp. 305-319.

[13] ———, *On estimating the "rates of convergence" of iterative methods for elliptic difference equations*, Trans. Amer. Math. Soc., **114** (1965), pp. 320-354.

[14] ———, *On the eigenvalues of second order elliptic difference operators*, SIAM J. Numer. Anal., **19** (1982), pp. 518-530.

[15] S. V. PARTER AND M. STEUERWALT, *Block iterative methods for elliptic and parabolic difference equations*, SIAM J. Numer. Anal., **19** (1982), pp. 1173-1195.

[16] J. RICE, *On the effectiveness of iteration for the Galerkin method equations*, in *Advances in Computer Methods for Partial Differential Equations IV*, R. VICHNEVETSKY AND R. S. STEPLEMAN, eds., IMACS, New Brunswick, NJ, 1981, pp. 68-73.

[17] D. J. ROSE AND R. A WILLOUGHBY, eds., *Sparse Matrices and Their Applications*, Plenum Press, New York, 1972.

[18] G. STRANG, *Approximation in the finite-element method*, Numer. Math., **19** (1972), pp. 81-98.

[19] G. STRANG AND G. J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[20] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[21] ———, *Extensions of the successive overrelaxation theory with applications to finite element approximations*, in *Topics in Numerical Analysis*, J. J. H. MILLER, ed., Academic Press, New York, 1973, pp. 329-343.

[22] D. M. YOUNG, *Iterative methods for solving partial difference equations of elliptic type*, Trans. Amer. Math. Soc., **76** (1954), pp. 92-111.

[23] _____, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

[24] O. C. ZIENKIEWICZ, *The finite element method: from intuition to generality*, Appl. Mech. Rev., **23** (1970), pp. 249-256.